Calhoun: The NPS Institutional Archive
DSpace Repository

Faculty and Researchers

Faculty and Researchers' Publications

2002-07

Virtual Multimedia Libraries Built from the Web

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

<https://hdl.handle.net/10945/36953>

Downloaded from NPS Archive: Calhoun

<http://www.nps.edu/library>

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

Virtual Multimedia Libraries Built from the Web

Neil C. Rowe

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

ABSTRACT

We have developed a tool MARIE-4 for building virtual libraries of multimedia (images, video, and audio) by automatically exploring (crawling) a specified subdomain of the World Wide Web to create an index based on caption keywords. Our approach uses carefully-researched criteria to identify and rate caption text, and employs both an expert system and a neural network. We have used it to create a keyword-based interface to nearly all nontrivial captioned publicly-accessible U.S. Navy images (667,573), video (8,290), and audio (2,499), called the Navy Virtual Multimedia Library (NAVMULIB).

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information Filtering.

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Multimedia, images, libraries, World Wide Web, captions, information retrieval, video, audio.

1. INTRODUCTION

The World Wide Web provides a vast number of images and increasing numbers of video and audio. Rather than copy those multimedia objects into digital libraries for easier access, it makes more sense to index the objects in situ to create virtual libraries. But this is harder than it seems because captions and descriptions are inconsistent and inconsistently placed, and content analysis of completely unconstrained multimedia is beyond the current state of the art. Just indexing multimedia through text search engines is quite imprecise; in a random sample we took, only 1.4% of the text on Web pages with images described those images. Commercial systems like AltaVista Image Search only index the easy-to-see image captions like text-replacement ("ALT") strings, achieving good precision (accuracy in the images they retrieve) but poor recall (thoroughness in finding relevant images). Effective indexing requires some judgment that recognizes a large number of clues. Some good work has been done on image indexing in the past using a limited set of clues and restricted kinds of images [1,3,4], but no one has attempted a truly comprehensive attack on the multimedia indexing problem.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Joint Conference on Digital Libraries '02, July 8-12, Portland, Oregon.

Copyright 2002 ACM 1-58113-000-0/00/0000...\$5.00.

2. THE SOFTWARE

MARIE-4 (see [2] for more details on the image-finding part) is written in Java. Its first phase is a "crawler" or "spider" that automatically searches part of the Web for caption candidates, given a starting page and the number of trailing domain words establishing locality (so "www.nps.navy.mil 2" indicates all "navy.mil" sites). Links are explored from the starting page in breadth-first search using order of discovery for links at the same depth. Link types extracted include straight HREF constructs, area and image maps, and Javascript constants. Duplicate links are ignored, and page content is hashed to enable ignoring of the same page stored under different link names.

Since multimedia descriptions appear in many forms, a good multimedia indexer needs to address a considerable range of clues. We distinguish clues shown in experiments to be at least 97% certain (e.g. "Figure 7") from clues merely suggestive (e.g. that the associated image is JPEG format rather than

GIF). We handle strong clues by a rule-based expert system (applied first to reduce data volume) and the suggestive clues by a neural network. The strong clues apply to a list of potential caption types that are searched for in the source text of the Web page. They include: the important words of the full multimedia link (URL); any text replacement for the multimedia ("alt"); clickable text that retrieves the multimedia; text delineated by HTML constructs for fonts, italics, boldface, centering, table cells and rows, and (rare) explicit captions; unterminated or unbegin

paragraph constructs; the title and nearest-above headings on the page (but not the unreliable "meta" keywords); and specific word patterns of multimedia-object reference (e.g. "in the photo above") found by partial parsing. Pruning then eliminates HTML and Javascript syntax from captions; too-small and too-narrow images (usually graphics); images appearing more than once on a page (usually control icons); and images appearing on three or more different pages. We then use the Web to query the status of the remaining multimedia objects to confirm they still exist and are indeed in an appropriate multimedia format.

We then rate captions by weighting suggestive clues. Nonlinear functions are applied to the factors to give consistent mean values and standard deviations. A small neural network is used with inputs derived from the individual words of caption, the destemmed words of the image filename, the intersection of the two, the type of caption, its length, and the image size. Weights are the conditional probabilities of a clue occurring in a caption when it occurs at all. Examples are the caption-word clues "during" (weight 0.79), "photograph" (0.73), "download" (0.03), and "update" (0.00); the image-filename word clues "people" (0.65), "page" (0.33), and "button" (0.33); and the caption-type clues "<a>" (for text displayed with a link, 0.66), "<h2>" (heading font, 0.46), "<td>" (table entry, 0.43), "<i>" (italics, 0.42), "alt" (alternative text for an image, 0.27), "<center>" (0.12), the image filename itself as a caption (0.09), and "<h4>" (small heading font, 0.00). Weights were estimated from a training set of 8140 candidates from Web pages found in "random walks" from a set of representative starting pages, of which 27% were captions (as confirmed on inspection by the author, with minimal bias since captions are generally obvious). We consider as significant only those clues whose weights are further than one standard deviation from the mean of a binomial distribution. A neural network is preferable here to Naïve-Bayes and association-rule methods as the clues are strongly correlated, preferable to decision trees since there are unlikely to be complex logical relationships between clues, and preferable to case-based reasoning since there are no "ideal" captions.

Caption words are then indexed, and the indexes used by a keyword-based graphical user interface running on a Web server. Users give keywords describing the multimedia they are looking for, and the indexes are used to locate and rate matching Web objects; the best matches are displayed (or if audio or video, pointers to them) together with all their inferred captions. Caption ratings from the neural network are scaled by additional factors determined from careful experiments: We weight higher the shorter captions containing a keyword, captions with the keywords occurring earlier in the caption, captions with similar capitalization as the keywords, and captions in which keywords are adjacent.

3. TESTING THE SOFTWARE

Our work has gone well beyond previous work in the way we have tested caption clues systematically for usefulness. Figure 1 shows an example, the recall-precision curves on a random test set of 3945 random captions and noncaptions for nine important clues. Not surprisingly, caption-word clues were the most helpful; other intuitive clues do not fare as might be hoped, or are useful only for a range of recall values. We use studies like this to decide which clues to include in caption weighting and to assign initial weights; weights are then iteratively optimized using steepest-descent methods. Some tests illustrated dangers of overtraining, as several locally optimal sets of weights for parts of the training set had poor performance on the remainder; this suggests that a good training set for this domain needs to be large. We recently demonstrated eight times the recall for the same level of precision as AltaVista Image Search.

Similar tests were done on properties of the images themselves: size, number of colors, count of the most-frequent color, average color saturation, and average color variation between neighboring pixels. The idea was that the larger or more varied images might more likely be captioned. But in contrast to our earlier research, only image size proved significant help in a random sample, as captionable images can vary considerably in appearance.

Other experiments confirmed that bootstrapping (learning from experience) could iteratively improve performance, by taking the words of the strongest caption candidates as new word clues for rating further pages. While not as good as using human-judged captions for extracting word clues (a 3% improvement versus a 10% improve in precision for recall below 50%), this self-improvement can provide the basis for indexing the entire Web effectively.

4. THE NAVY VIRTUAL MULTIMEDIA LIBRARY

Our most comprehensive test was construction of an index to all the images, audio, and video on "navy.mil" Web sites in December 2001. We deliberately chose humble hardware, a single PC connected through our school's busy Internet interface, to demonstrate the ease of the approach. It did take 35 days of real time to process the requisite 574,887 Web pages. 2,193,792

Figure 1: Precision vs. recall for various caption clues.

image-caption and 35,770 audio-caption and video-caption candidates were found for respectively 667,573 distinct images, 8,290 distinct video clips, and 2,499 distinct audio clips. 1,814 had at least one indexable audio or video clip. 680,022 additional captions were eliminated because their images appeared on three or more pages, and 3,123,724 because either their images no longer existed after a month or were found to be too small. We provide public keyword-based interfaces using Java servlet software, and provide a page of all caption-media links for audio and video. A typical three-word query needs around 10 to 90 seconds for the answer. We automatically collect usage statistics on the browsers to provide study opportunities. Public access to NAVMULIB is at <http://triton.cs.nps.navy.mil:8080/rowe/navmulib.html>.

5. REFERENCES

- [1] Mukherjea, S. and Cho, J. Automatically determining semantics for World Wide Web multimedia information retrieval. *Journal of Visual Languages and Computing*, 10 (1999), 585-606.
- [2] Rowe, N. A high-recall self-improving Web crawler that finds images using captions. To appear in *IEEE Intelligent Systems*, late 2002; also at www.cs.nps.navy.mil/people/faculty/rowe/crawl.htm.
- [3] Sclaroff, S., La Cascia, M., Sethi, S., and Taycher, L. Unifying textual and visual cues for content-based image retrieval on the World Wide Web. *Computer Vision and Image Understanding*, 75, 1/2 (July/August 1999), 86-98.
- [4] Srihari, R. Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review*, 8, 5-6 (1995), 409-430.