

# Increasing Web Accessibility by Automatically Judging Alternative Text Quality

Jeffrey P. Bigham

Department of Computer Science and Engineering  
University of Washington  
Seattle, WA 98195 USA  
jbigham@cs.washington.edu

## ABSTRACT

The lack of appropriate alternative text for web images remains a problem for blind users and others accessing the web with non-visual interfaces. The content contained within web images is vital for understanding many web sites but the majority are assigned either inaccurate alternative text or none at all. The capability to automatically judge the quality of alternative text has the promise to dramatically improve the accessibility of the web by bringing intelligence to three categories of interfaces: tools that help web authors verify that they have provided adequate alternative text for web images, systems that automatically produce and insert alternative text for web images, and screen reading software. In this paper we describe a classifier capable of measuring the quality of alternative text given only a few labeled training examples by automatically considering the image context.

**ACM Classification:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - *Information filtering*.

**General terms:** Design, Human Factors

**Keywords:** web accessibility, alternative text

## INTRODUCTION AND BACKGROUND

Blind users employ screen readers as their interface to the information on the web and are currently at a disadvantage because alternatives are not adequately provided for visual content. When a person using a screen reader reaches a web image while reading a page, the screen reader substitutes alternative text provided for the image in place of the visual information they cannot see. If the web author provided no alternative text, then the screen reader will either read the filename of the image or read nothing, depending on the user's preferences. If the image contained important information, then this information is inaccessible to the user. Alternatively, many images serve only a visual purpose, such as those images used for decorative or structural

purposes. These images should be assigned a zero-length alternative text so the screen reader knows that they should be skipped. Often they are provided alternative text and this alternative text is rarely helpful, which both annoys users and causes them to be less efficient. These images are often assigned long alternative text, such as "1x1 clear pixel," and exist many times on the web page.

Of the images used to convey important page content, nearly 40% are not assigned alternative text[2]. Even those images that are assigned alternative text are often assigned inaccurate alternative text to the frustration of users[8]. A U.S. District Court recently found that Target could be sued under the Americans with Disabilities Act (ADA) because of the inaccessibility of its web site[1]. A main complaint of the plaintiffs was the lack of alternative text for vital images on Target's web site. In light of this case and the impact of the continuing inaccessibility of the web, it is important to make proper web accessibility as easy to achieve as possible for those that can influence it.

Tools that verify the accessibility of web sites currently ignore the quality of alternative text because no automatic method of judging the quality of alternative text is available. Other approaches to solving the problem of web images with poor or missing alternative text concentrate on the client interface. For example, Bigham *et al.* recently introduced WebInSight, a system that automatically formulates and inserts alternative text for many web images as users browse the web[2]. The system is necessarily conservative in its generation of alternative text because it does not have the ability to judge the likelihood that arbitrary alternative text will be helpful to users. To overcome the annoyance of inappropriate alternative text, many screen reader users write custom scripts in their screen reading software that prevents alternative text that matches alternative text that is known to be bad, such as "image," "spacer," or "\*", from being spoken. To create such rules the user must be technically savvy and the rules created are limited to what the user has encountered.

All of these tools would be much better equipped to increase web accessibility if they could automatically judge the quality of alternative text. An automatic classifier that accurately judges the quality of alternative text stands in opposition to the commonly held belief that the appropriateness of alternative text is too subjective to be determined automatically. Judging alternative text quality is similar to the task of decid-

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IUI'07, January 28–31, 2007, Honolulu, Hawaii, USA..

Copyright 2007 ACM 1-59593-481-2/07/0001 ...\$5.00.

ing upon accurate keywords for image search (another difficult problem) but differs in two ways. First, the images that most need alternative text are those that lack the easiest sources of accurate keywords (such as the `alt` attribute of the `img` tag). Second, the appropriateness of alternative text is highly dependent on the context where the image appears. Luis von Ahn *et al.* has introduced clever games designed to entice humans to label images[12, 11]. These games entice users to provide labels with neither knowledge that their labels will be used as alternative text nor of an image's original context on the web. While labels produced by such games may potentially form valuable alternative text, the quality of alternative text produced in this way needs to be evaluated.

In this paper, we first describe a classifier that is capable of differentiating between good and bad alternative text. We introduce a number of interesting features used in the construction of the classifier that we argue are informative in deciding the appropriateness of alternative text for a particular image. We then describe preliminary results of tests we conducted with our classifier that show our classifier performs well in classifying alternative text found on the web.

## THE CLASSIFIER

In order to judge the quality of alternative text, we built a classifier that uses features based primarily on the alternative text content and on the context of where it is located on the web. The system takes as input an arbitrary list of URLs and, relying on a few simple rules, is capable of bootstrapping an initial classifier. The classification problem explored here is similar in many ways to the problem of web page classification which has been explored extensively[9, 3]. Marques *et al.* developed a tool that used machine learning techniques to aid in the semantic markup of images, but this tool assumed that labels were not already given and that the purpose of labeling the images was for the semantic web[7].

Our classification problem is to judge if arbitrary alternative text is more likely to be appropriate or inappropriate given its context and the use of the alternative text as an equivalent substitute for the image. Using this information, users could decide to suppress alternative text likely to be inappropriate and web designers could take a second look to ensure that any alternative text that the system finds inappropriate is truly equivalent to the visual content of the image.

## FEATURES

Our classifier relies upon a number of features derived from the surrounding context. Each feature attempts to measure one of the following: the similarity of the alternative text to the content of the page where it was found, how similar the alternative text is to alternative text that is known to be good or bad, how similar the alternative text is to other methods of labeling the image, the likelihood that the alternative text is composed of content with meaning, and features of the image that it is describing. These features are listed individually in Table 1 and the motivation for each discussed below.

### Content Similarity

The first set of features is based on the similarity between alternative text content and the content on the web where the alternative text is found. We expect alternative text that is

closely related to the content of the web page where it is found to be informative with respect to the images on the page. Likewise, we expect alternative text that is unrelated to the content on the page to be inappropriate.

For instance, the text "Click Here" is a commonly used for alternative text, but, unless the image contains the text "Click Here," this alternative text neither conveys the content of the image nor what will happen if the image is clicked. This text is unlikely to be more common on a particular web site than it is elsewhere on the web. Similarly, the alternative text "empty chessboard" is more likely on a site whose content centers around chess than it is on the rest of the web, and, therefore, we might be more likely to believe this to be appropriate alternative text when it is found on such a site. From this intuition, we include a Boolean feature computed for the alternative text  $A$  at the domain *www.domain.com* using the following inequality and the Google Search Engine:

$$\frac{\text{hits}(A \text{ site:}www.domain.com)}{\text{hits(site:}www.domain.com)} > \frac{\text{hits}(A)}{\text{\# pages in Google index}}$$

This feature was inspired by PMI-IR, a metric originally introduced in the context of automatically solving synonym test questions[10].

### Word-based Similarity

As mentioned earlier, users of screen readers often write scripts that suppress alternative text from being read if it matches alternative text known to be unhelpful. Craven *et al.* found that five of the most common alternative texts assigned to images are unhelpful, including "1", "pad" and "\*" [4]. Other words are unlikely to be used as generic substitutes in this fashion (for example, "people," "driving," or "ice hockey"). One of the features used in our classifier is the output of a Naive Bayes text classifier that uses a bag-of-words model to learn the words that distinguish appropriate alternative text from inappropriate alternative text. Similar models have been used for text classification before[3, 6].

### Automatic Labeling Similarity

Recent systems have attempted to automatically derive alternative text for web images and can do so quite accurately for many images. For instance, WebInSight offers two fully automatic labeling methods: link-based labeling and optical character recognition (OCR) labeling[2]. The link-based labeling method provides alternative text for images that serve as links. To formulate alternative text for such an image, the link is followed and candidate alternative text is returned from the title and headings present on the linked page. The OCR labeling module is designed to provide alternative text for images that contain text. It implements a series of preprocessing steps that prepares web images for OCR processing. When these labeling methods apply, they are quite accurate, and, therefore, comparing the content of alternative text of unknown quality to the alternative text provided by these methods can be quite informative as to its appropriateness.

Web authors commonly assign alternative text to images based on the filename that happens to be assigned to the image. While occasionally this can be an acceptable strategy if the image was assigned a meaningful name, more often this results in meaningless alternative text. For example, an im-



|  |   |   |
|--|---|---|
| <b>Examples</b>  |  |  |
| Alternative Text   | CS&E logo   | center  |
| Image Filename   | cse_logo_80x133.gif   | centerpict.jpg  |
| Found on Web Page  | www.cs.washington.edu   | www.cs.ucla.edu   |
| Appropriate Alt. Text                                    | Yes   | No  |
| <b>Features</b>  |   |   |
| Alternative text more likely on this site than on web.   | true  | false   |
| Vector cosine of alt. text to context and OCR alt. text. | 0.50  | 0.0   |
| Vector cosine of alt. text to filename-based alt. text.  | 0.41  | 0.50  |
| Alternative text exists on web page.                     | false   | true  |
| % of words in alt. text in web page.                     | 0.5   | 0.0   |
| % of words in dictionary.                                | 0.50  | 1.00  |
| Alternative text phrase in Google.                       | true  | true  |
| Number of words in alternative text.                     | 2   | 1   |
| Naive Bayes classification probability.                  | –   | –   |

Table 1: The features used for classification of alternative text, along with the values of these features for several examples. The values for features calculated only with respect to a given data set are omitted.

age with the filename `image45302.jpg` may be assigned the uninformative alternative text “image45302.” Filename analysis produces labels by splitting the assigned filename of a web image into words and potential phrases based on simple heuristics that break filenames into words by replacing characters commonly used to separate words (such as underscores and dashes) and by doing a dictionary-based search for words within the filename. Alternative text of unknown quality can be compared to the alternative text produced by this method to gain insight as to its appropriateness.

### Meaningful Alternative Text

Valuable alternative text is usually composed of words and phrases that can be independently verified to be meaningful. If the individual words in alternative text exist in a dictionary or on the web, then they could form part of appropriate alternative text. The arbitrary strings, sequences of punctuation and complex filenames that are often used incorrectly as alternative text won’t be found in a dictionary. Both the fraction of words in an alternative text string that are found in either the dictionary or on the web and a Boolean value indicating whether the entire alternative text string is found in the Google search engine are used as features in our classifier.

### EVALUATION

For our evaluation, we collected two sets of training data. Data set *A* consisted of 250 images with non-null alternative text sampled from 100 web pages chosen randomly using the Yahoo Search Engine. For each image in data set *A*, both the alternative text provided originally and the alternative text calculated by filename analysis were classified manually as either good or bad. Data set *B* consisted of all 8350 images with non-null alternative text on the 248 high-traffic English-language web pages considered in the image accessibility studies presented by Bigham *et al.*[2]. This data set was left unlabeled.

In our first experiment, we trained an AdaBoost classifier[5] using the labeled examples of originally provided alternative text in data set *A*. We performed 5-fold cross-validation anal-

ysis of our classifier on these examples and found that our classifier labeled the examples in the data set with an accuracy of 86.8%. We then repeated this experiment using the examples of alternative text provided by filename analysis and found that the classifier labeled these examples with an accuracy of 77.6%. While 60.4% of the alternative text produced by filename analysis was incorrect originally, using the classifier to filter out the alternative text that the classifier believes to be bad reduced the percentage of bad alternative text to 8.8%. Filename analysis became much more usable because it could correctly label 36.2% of images, and this only cost 13.6% of the good alternative text being suppressed.

Next, to explore how more examples might be added without additional manual labeling, we boot-strapped classifications for the examples in set *B* and evaluated a classifier trained on these examples on the labeled examples in data set *A*. To label the examples in data set *B*, we did not explicitly analyze the appropriateness of the alternative text provided. Instead, we leveraged the empirical observation that the extent to which a web page follows accessibility guidelines often determines the appropriateness of the alternative text contained on it. In this case, examples of bad alternative text are drawn from all alternative text provided for images that should be left unlabeled. Examples of good alternative text are drawn from all alternative text provided for images on pages that label all images that should be labeled and assign all images that shouldn’t be labeled a zero-length alternative text. To determine if an image should be labeled, we used a heuristic introduced by Bigham *et al.*[2]. The heuristic states that an image should be labeled if it is multi-colored and larger than 10 pixels in both dimensions or if it has an associated function (such as serving as a link).

After selecting alternative text using this process, 239 images were classified as having good alternative text and 112 as having bad. Unfortunately, we could not use such automatic labeling to judge the quality of all alternative text because it fails to label 7999 of the images, including many exam-

ples of bad alternative text that were assigned to images that needed labels. We trained an AdaBoost classifier[5] using the 351 total examples labeled in this way and tested it by labeling the examples in data set *B*. The classifier was able to correctly label 86.3% of these examples, which suggests that classifiers might benefit by automatically incorporating additional data in this way.

## CONCLUSION AND FUTURE WORK

Our preliminary results indicate that it is possible to automatically judge the quality of alternative text. Our features seem to capture valuable characteristics of alternative text quality, although more analysis is warranted and planned. Our tests were carried out on a small data set and we hope that by expanding it we can gain a deeper understanding of the potential of judging alternative text automatically.

An interesting way to gather large amounts of higher-quality data would be to integrate user feedback into a screen reader, allowing users to indicate the quality of alternative text that they encounter. The approach presented in this paper does not allow for the complex interactions that most likely govern whether arbitrary alternative text is appropriate or not. We suspect that web authors provide inappropriate alternative text for a number of underlying reasons, and modeling them more directly might both suggest improvements to our metric or provide insights into why web designers don't always provide adequate alternative text in the first place.

Our goal in developing a method for automatically judging the quality of alternative text is to improve systems that currently deal with alternative text. We plan to use our metric in a tool for web designers that gives them feedback on the quality of the alternative text they provide, as opposed to current tools that can only verify the existence of alternative text. While automatic methods for determining the quality of alternative text will never be as accurate as trained humans, highlighting those images that are assigned alternative text of questionable quality will allow humans relatively unskilled in web accessibility to produce more accessible web pages. An important test for such a tool, beyond simply the ability to detect images that have not been assigned appropriate alternative text, will be its ability to affect the accessibility of web pages produced using it.

Finally, we plan to integrate our metric into our WebInSight system to help the system more accurately choose which of many potential labels to provide as alternative text to users. Another possibility, is to explore the effect of integrating the metric into screen reading products. Based on the fact that many users choose to hide certain alternative text, it seems likely that they would appreciate the option to hide alternative text that is likely to be inappropriate and uninformative. In this situation false positives in particular will be most costly because users often prefer the annoyance of unhelpful information to losing information of value.

## ACKNOWLEDGMENTS

This research was funded by National Science Foundation grant IIS-0415273.

## REFERENCES

1. National Federation of the Blind vs. Target Corporation. U.S. District Court: Northern District of California, 2006. No. C 06-01802 MHP.
2. J. P. Bigham, R. S. Kaminsky, R. E. Ladner, O. M. Danielsson, and G. L. Hempton. Websight: Making web images accessible. In *Proceedings of Eighth International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '06)*, October 2006.
3. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory (COLT '98)*, pages 92–100, New York, NY, USA, 1998. ACM Press.
4. T. C. Craven. Some features of alt text associated with images in web pages. *Information Research*, 11, 2006.
5. Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, 1996.
6. D. D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
7. O. Marques and N. Barman. Semi-automatic semantic annotation of images using machine learning techniques. In *International Semantic Web Conference (ISWC)*, Sanibel Island, FL, USA, 2003.
8. H. Petrie, C. Harrison, and S. Dev. Describing images on the web: a survey of current practice and prospects for the future. In *Proceedings of Human Computer Interaction International (HCII) 2005*, July 2005.
9. A. Sun, E. Lim, and W. Ng. Web classification using support vector machine. In *WIDM '02: Proceedings of the 4th international workshop on Web information and data management*, pages 96–99, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-593-9.
10. P. Turney, M. Littman, J. P. Bigham, and V. Shnyder. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 482–489, 2003.
11. L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of Computer Human Interaction (CHI) 2004*, April 2004.
12. L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 79–82, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-372-7.