Name That Link: Intelligent Link Anchor Generation

Steve Martin[1]
University of Washington, Seattle, WA, USA
stevaroo@cs.washington.edu

Technical Report 02-10-02
October 2002

Abstract

A useful and growing web technology is the ability to present dynamically customized web content to users on the fly. Customized content often includes customized links, so a natural question arises: what should the anchor text be for dynamically generated links?

This paper addresses the question of how to label dynamically generated links informatively and understandably. Our approach uses heuristics to generate many possible anchor labels, evaluates these labels, and then chooses the best anchor text for the link.

1 Introduction

Many sites on the Web today perform some kind of dynamic content customization for their visitors. How this optimization is done varies, but often the goal is to improve site navigation for users. For example Proteus [ADW01a] re-structures web pages in a manner to improve their display on hand-held devices. In e-commerce, Amazon.com dynamically adds links to books, etc, that the visitor may want to buy based on past behavior and purchases by other visitors.

A significant problem in dynamically adding links is labeling them. Informative anchor texts are important in that they, alone, describe for users the destination of the link. In the general case, a site might also want to dynamically link to other domains on the web, so simply recording metadata for pages within the same domain is insufficient.

One possible approach to this problem is to automatically summarize the destination page and use this summary text as the anchor for the link. However, web pages vary tremendously: some pages contain only text, others combine text and images, while still others display only with third-party plug-ins. In addition, although standards exist, not very many sites fully comply with them. Summarizing text, itself, remains an open AI challenge; these web-specific factors combine to make it even more difficult.

A second approach is to use the contents of the <title> tag from the destination page as the anchor text of the link to that page. This approach is taken by most search engines. Unfortunately, using the title of a page makes the link text entirely dependent on the author of that page. If the title of the page is missing, overly verbose, or non-

1. Also contributing were Corin R. Anderson (corin@cs.washington.edu) and Lesley Forbush (lesley@cs.washington.edu)

informative, (e.g., consider a page entitled "Homework" at UW/CSE, or "News:World:Middle East:Isreal:2001:June:6 on SeattleTimes.com) then the link text will be poor.

Our approach is to generate candidate anchor texts, evaluate these anchor texts, and then choose the best one for the link. In the first step, we use heuristics to process a page and return candidate anchor texts. Next, we use a decision tree to evaluate these anchor texts. The decision tree is trained using human input on anchor texts generated by our heuristics as examples. Finally, the label that is rated the highest is then used as the link text.

2 Generating Possible Anchor Texts

Anchor texts should concisely and accurately describe the content of the destination page. To this end, we present three heuristic functions that produce possible anchor texts that generally meet these conditions.

2.1 Heuristic 1: Using the title of a page

Our first heuristic uses the title of the page as the single line description of that page. This approach is used by search engines, such as http://www.google.com, when summarizing search results.

This idea works moderately well as long as the page passed to the heuristic has a meaningful title. However, page authors will often fail to provide useful titles—or even any title at all. **Figure 1** shows the user ratings of page titles as anchor texts based on a study link we conducted. (See **Section 3** for more details about the user study.)

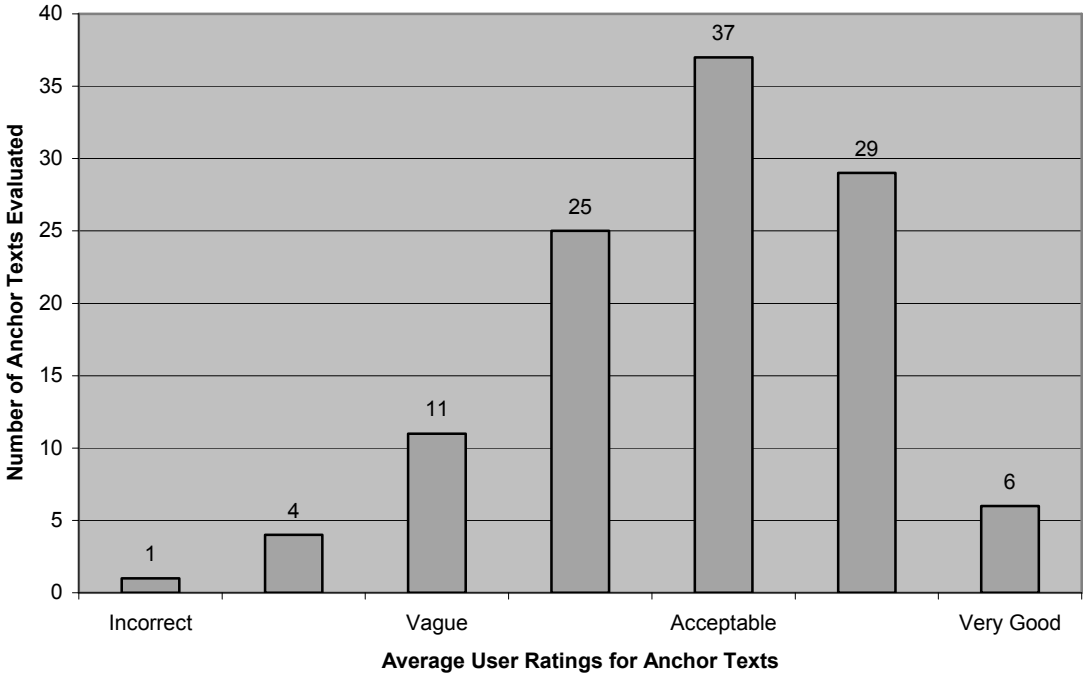

Figure 1. Distribution of user ratings for anchor texts consisting of page titles.

As **Figure 1** demonstrates, page titles typically work well as link labels—human viewers rated 64% of the titles generated using this method as "Acceptable" or better—but are often inadequate. For example, member pages that use frames share the title of the frameset document; these pages may be very diverse. Also, some pages are display content using third party media plug-ins, such as Macromedia Flash. We found these pages seldom make use of a title tag. Finally, many pages just don't have meaningful page titles.

2.2 Heuristic 2: Using link labels from other pages

Our second heuristic uses the anchor text descriptions of a given page from other pages to generate a link label for that page. The overall goal is to exploit the context the destination page is presented in on other pages as possible link texts. **Figure 2** illustrates this idea.

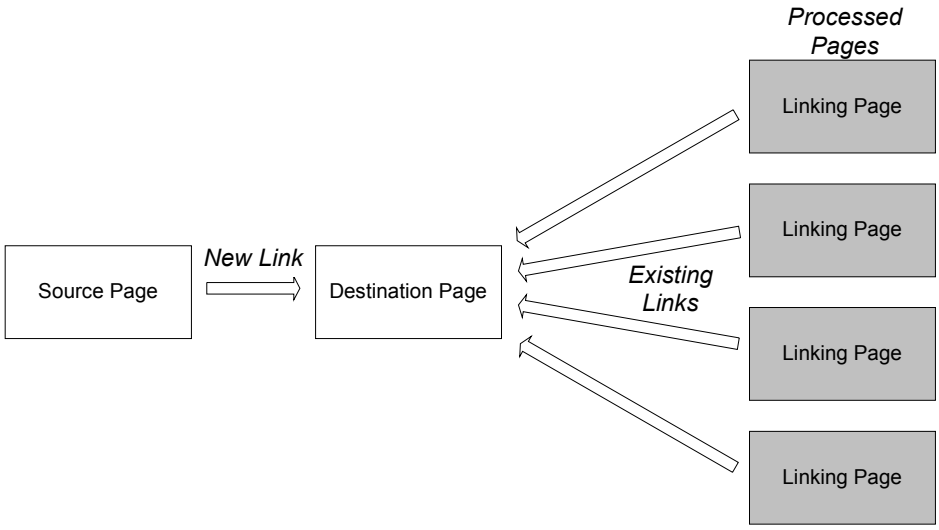

Figure 2. Description of the process the heuristic uses to generate a link text. The link from the source to the destination is generated by examining the anchor labels of *other* pages linking to the destination page.

Our heuristic uses Google to find pages that link to a given destination page, by encoding a special URL string for Google instructing it to find *m* pages linked to the destination page. A proprietary tool could be easily substituted for this step; we chose Google for convenience. The result page returned by Google contains links to each of the *m* pages. We found that a good value for *m* is around 20. Each page in the list of linking pages is then parsed for the label of the link to the destination page. These descriptions are then sorted by number of occurrences, and the top *n* candidates are returned. In our tests, *n* was set to 5. **Figure 3** displays the results of a user study evaluating the utility of this heuristic.

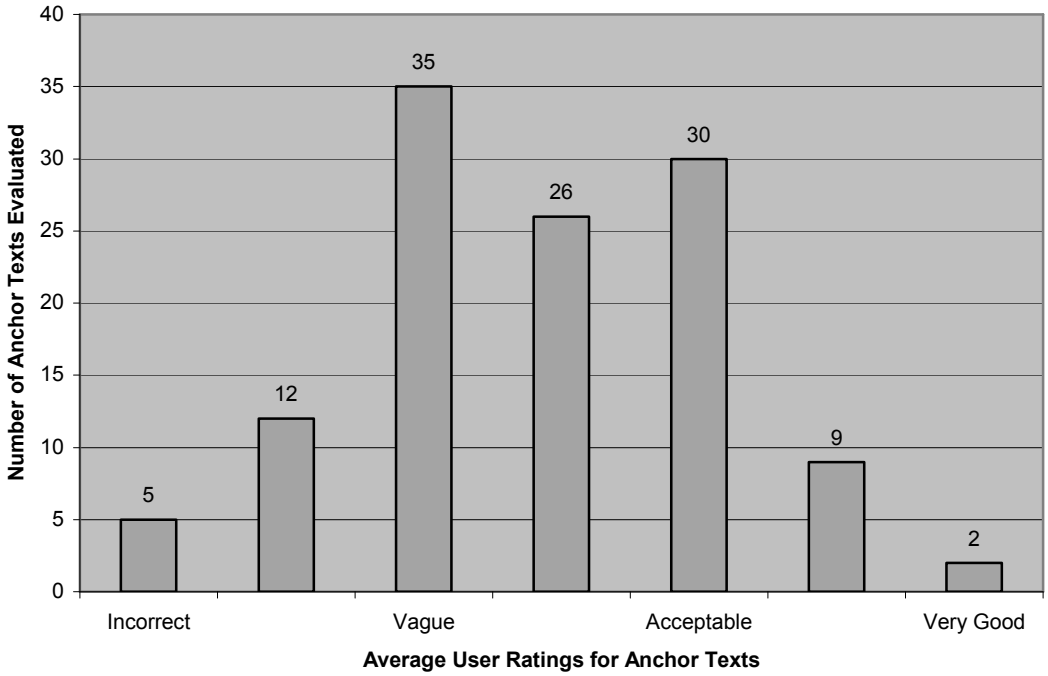

Figure 3. Distribution of user ratings for anchor texts consisting anchor labels of links to this page from other pages.

As **Figure 3** demonstrates, this method of generating anchor texts does not work quite as well by itself as the first method did. However, not all pages have others that link to them. In particular, many pages, such as news articles, are so specific that no page outside of their immediate parents links to them. In addition, not all link texts are descriptive outside of the context of their respective pages. As an example, links that aren't text at all, such as pictures or movies, will not give any anchor text.

This heuristic is useful because it will work whenever there are any back links to the destination page. For this reason, it can return a good link text even when the title of the page is nondescript or overly verbose.

2.3 Heuristic 3: TFIDF

Our third heuristic looks at the actual words themselves on the source and destination page and picks out the most important phrases that differentiate the pages from each other. This method has two parts: ranking individual words based on their value as part of the final link text, and ranking phrases to determine the best anchor label.

The heuristic begins by generating lists of word stems, one list for each of the source and destination pages. A word stem is a linguistic abstraction that describes the root of a given word. For example, *child* is the stem of *children*, and *happi* is the stem of *happy* and *happiness*. We map the real words into their stems because we are concerned only with the concept of each word, not the exact form itself. The number of occurrences of a stem word on each page is considered its rank. Words that are known not to be

descriptive, such as *a*, *is*, and *the*, are removed from the lists; these are called 'stop words.'

Next, the stem list for the destination page—the page we are interested in describing—is compared to the list for the source page. If a stem is in both lists, its rank in the destination page list is divided in half. This procedure follows the Term Frequency Inverse Document Frequency (TFIDF) approach in that the rank of a word on a given page is highest when it appears many times on that page but on few other pages. [SM83]

We observe that words that appear many times on a page are often not describing the page itself, but more likely jargon or highly specific but uninformative words, such as "cents" on an eBay page. Thus, the heuristic removes words from the destination page list that occur more than *m* times. We found no value of *m* that was conclusively best; values of 10, 100, and 10,000 all worked well. The resulting top ranking words from this list are then considered in the next segment of this heuristic.

The next step is to generate a list of phrases from the destination page. The heuristic considers the text contained in any HTML tag. For each of these possible phrases, the heuristic sums the rank of each word in the phrase and divides by the number of words. Any phrase longer than *o* characters is discarded; we found a good value of *o* to be 50 characters, as long phrases typically make poor link names. The score is doubled if the phrase is contained in a title, heading, or strong HTML tag. The highest scoring phrases are then returned as candidate anchor texts.

This method generates interesting candidates, but did not fare as well in our user study as the other two heuristics. **Figure 4** gives a visual distribution of how users ranked the link texts generated with this heuristic.

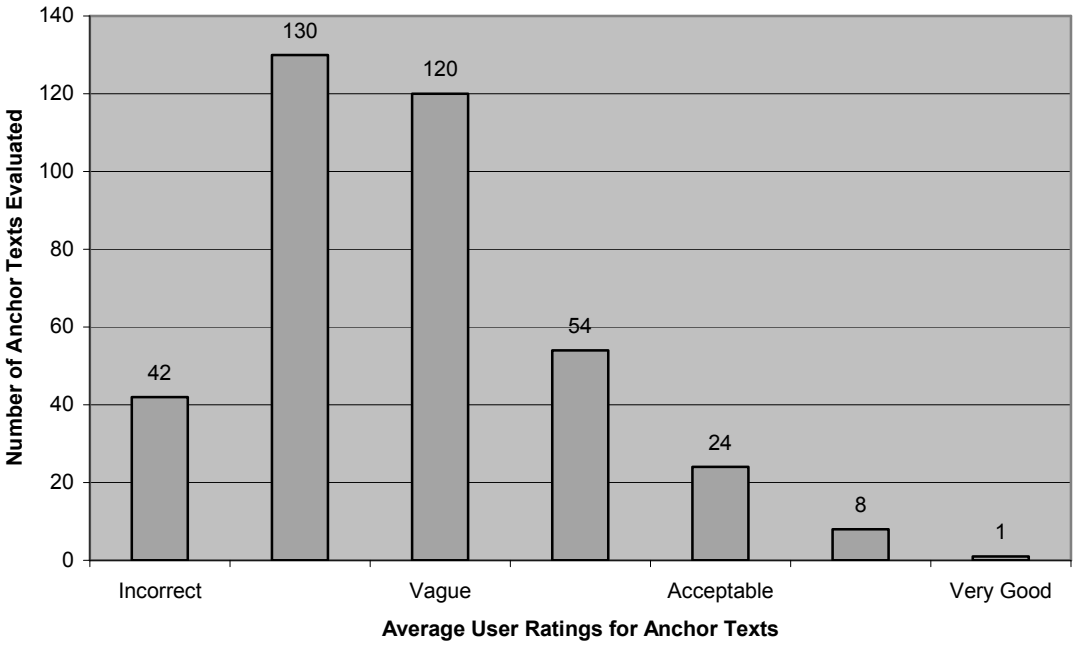

Figure 4. Distribution of user ratings for anchor texts generated using the TFIDF method.

Figure 4 shows that anchor texts generated using the TFIDF method have a fairly low approval rating from users. This is because the effectiveness of this heuristic is very dependent on the type of page it is used on. When this heuristic is used on pages with lots of repetitive text, such as classified listings, the anchor texts it returns are not very descriptive. Unfortunately, the page pairs used in our user study included many pages that are less than optimal for TFIDF to use, resulting in lower scores from our user study. However, when presented with a text-heavy page dealing with a single topic, such as a news article, TFIDF will give good results.

Unlike the previous heuristic, TFIDF can be effective on pages that do not have a large number of other pages linking to it. In addition, it can also return good anchor texts on pages that have non-descriptive or overly verbose titles. Thus, if the previous heuristics fail to return good anchor labels to a given page, TFIDF can possibly pick up the slack.

3 Evaluating Anchor Texts

As the previous section demonstrates, no single heuristic will always generate the best text for a link. We now present a combined approach, which uses all of our heuristics to generate candidate label; in this section we present an evaluation function.

The overall goal of this project is to generate link texts that are useful to human viewers, so the evaluation function must map to human approval. In order to learn this function, we performed a user study to gather data on how humans rate anchor texts.

The foundation for this study consists of several dozen pairs of source and destination web pages of our choosing (see **Appendix A**). These page pairs represent the source and destination of a link we need to label, and are classified into two major classes: links within a given domain (e.g. both pages within www.seattletimes.com), and links outside the domain. Study participants are then presented a randomized series of page pairs and a selection of possible link texts, generated using the heuristics outlined above, and asked to evaluate the accuracy/descriptiveness of each link text. A round of testing represented one user rating 20 source and destination pairs. Different rounds are generated using a different random seed, resulting in a pseudo-unique series of pages. **Figure 5** displays the layout of a typical question during the user study.

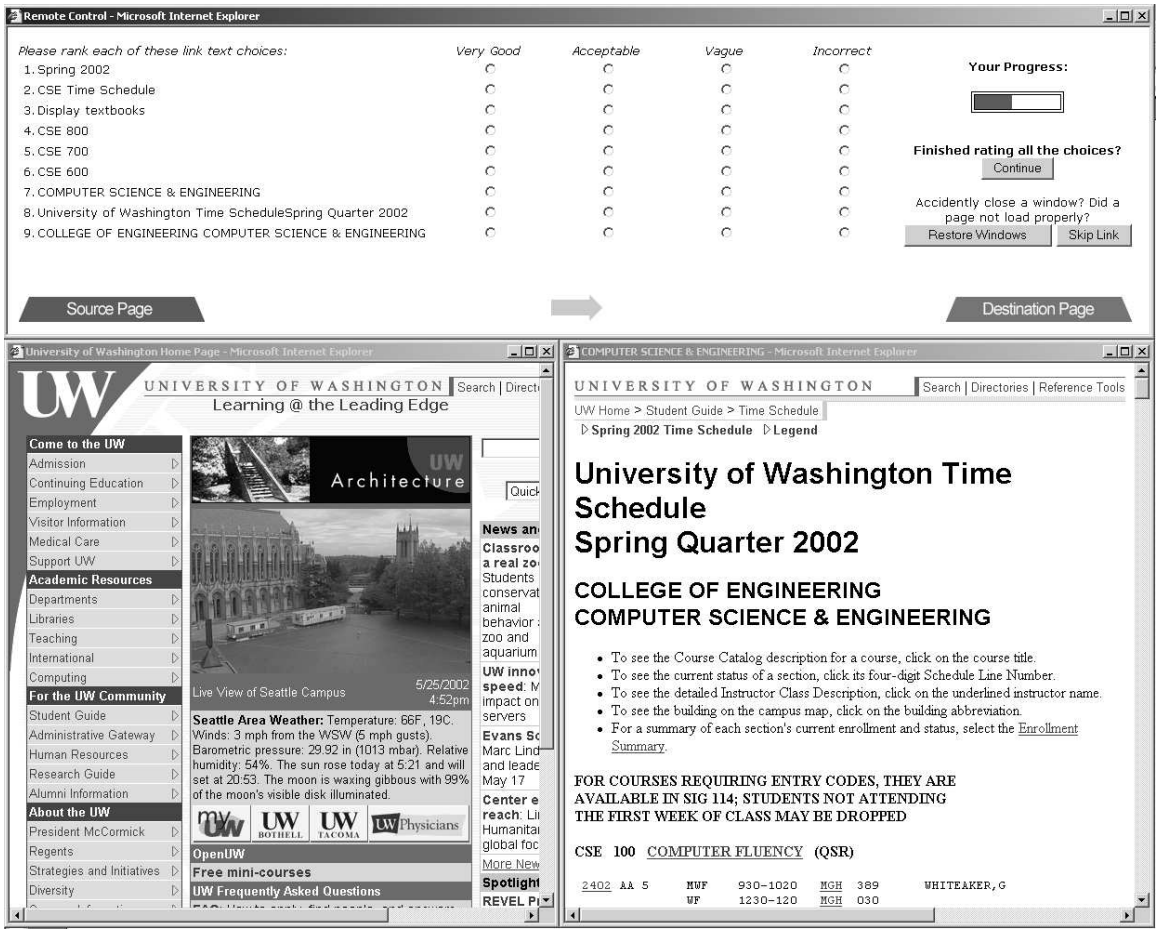

Figure 5. Screenshot of user study layout.

Respondents for the user study came from a limited pool of friends, family, and coworkers with varying degrees of Internet experience. The study lasted one week, and had 29 participants, yielding 36 rounds (including incomplete rounds) with 531 source and destination pairs reviewed. This resulted in 601 unique source, destination, and anchor text triplets.

With evaluation data gathered we trained a decision tree to predict the anchor label's human-judged rating. We used the UWML interface developed by Geoff Ghulten with C4.5 building the actual decision tree. [GH02]

When building the decision tree, we need to choose features to describe the unique triplet of anchor text, source page, and destination page. We chose:

- Whether each heuristic generated the text (i.e. separate features).
- The length of the source and destination web addresses.
- The length of the anchor text itself in words.
- The length of the anchor text itself in characters.
- Whether or not the source and destination are within the same or different domains.

- Whether or not the destination page has a title.

The class of each anchor text is the average score given to this link text by users. A rating of 'Incorrect" corresponds to 0, "Vague" to 1, "Acceptable" to 2, and "Very Good" to 3. Scores for each anchor text were averaged and rounded to the nearest 0.5.

4 Results

To complete our combined heuristic and decision tree approach, we randomize the data generated by the user study and build the tree. To make good use of all of the data, we perform ten-fold cross validation on the dataset with ten percent of the data forming our evaluation set. **Table 1** gives the average accuracy statistics generated by our decision tree.

Table 1. Accuracy statistics of our C4.5 decision tree using ten-fold cross validation.

Percentage of Completely Correct Classifications	31.0%
Percentage of Wrong Classifications	69.0%
Average Error Distance in Classifications	8.23
Standard Deviation of Error Distance in Classifications	6.18

As **Table 1** shows, the decision tree is completely accurate 31% of the time. Overall, the decision tree misclassifies the anchor text by slightly over one class on average. The accuracy of the decision tree is important because we want it to always choose the link that is best for human viewers.

To compare the combined approach to using any single heuristic, **Figure 6** shows the percentage of time each heuristic generated the top-scoring link text during the user study, and compares this value to the combined approach.

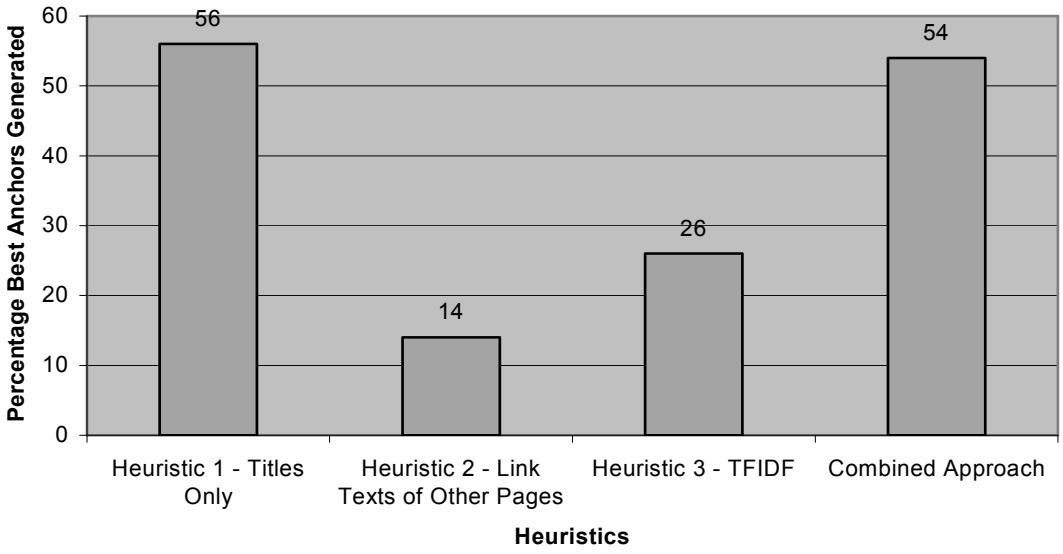

Figure 6. Percentage of anchor texts generated by each heuristic and the combined approach that are rated by human viewers as the best link text for that pair of pages.

As the figure demonstrates, each heuristic is returns the highest scoring link a significant percentage of the time. The combined approach returns the best link as decided by ratings given by human viewers 54% of the time. However, the title heuristic returned the highest scoring anchor text 56% of the time. There are several possible reasons for the relatively low score of the combined approach; the amount of data available to train the decision tree is somewhat small. In addition, there are only three heuristics to choose from. Finally, it's also possible that the attributes chosen to construct the decision tree were not optimal.

4.1 Future Work

The next step in generating more meaningful results would be to run a new user study. This time, users would be asked to evaluate link texts generated by a complete system that generates anchor labels from source and destination pairs using the combined heuristic/decision tree approach. These results could then be compared with the data already gathered to show that our approach to dynamically generating link texts is a solid one. However, while such a system does exist, time constraints prevent us from running such a study.

Other work that would improve results would be further research into better heuristics that deal with more of the web. The number of ways content is presented on the Internet is growing; none of the heuristics described thus far deal with multimedia such as pictures, movies, java applets, flash/shockwave content, etc. Unfortunately, multimedia is playing an increasingly important role on the web and should be considered in the scope of this research.

Once additional heuristics have been developed, it would be instructive to run a new user study with a wider sample of source and destination pages using all of the heuristics to generate a fresh data set. This data set could be analyzed with the process described above to provide fresh insight into our approach.

6 Conclusions

The number of pages presenting dynamic content is growing. However, when customizing a site for a specific user, the problem of labeling generated links arises.

There are a variety of ways to create informative link texts. One approach is to just use the title of the destination page. Another is to use the anchor labels of pages linking to the destination page. Finally, the text of the page can be summarized into a link text using methods such as TFIDF.

Our approach to this problem is to use a combination of these heuristics to generate a large pool of possible anchor texts for a link to a given page. From this pool, an evaluation function—implemented by a decision tree trained by user data—picks the best label for the link.

Analysis of our approach shows that this idea works: it generated the highest scoring anchor text (as rated by human viewers) more than half the time. However, when compared to using the heuristics by themselves, links generated from the title of the destination page had a higher percentage of the highest scoring anchor texts.

There are many possible improvements to the approach discussed in this paper. Other heuristics should be considered. Larger, more detailed user studies would be helpful to generate more specific data. Finally, further work might be done with a complete system implementing the combined method to see how well this idea works in the field.

References

[ADW01a] Corin R. Anderson, Pedro Domingos, and Daniel S. Weld. Adaptive Web Navigation for Wireless Devices. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*. 2001.

[ADW01b] Corin R. Anderson, Pedro Domingos, and Daniel S. Weld. Web Site Personalizers for Mobile Devices. In the *IJCAI Workshop on Inteligent Techniques for Web Personalization (ITWP)*. 2001

[GH02] Geoff Hulten. UWML. Private Communication. 2002.

[SM83] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, NY, 1983.