Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment

Barbara H. Kwaśnik, Kevin Crowston, Joseph Rubleske, You-Lee Chun

Contact Person: Barbara H. Kwaśnik
Hinds Hall, School of Information Studies
Syracuse University, Syracuse NY 13244, USA
bkwasnik@syr.edu
315 443-4547

This paper reports on one phase of a larger study whose overarching aim is to determine how providing genre metadata can help in access to sources of information in a digital environment. Of interest to us at this part of our study is the presentation of results once a search has been carried out. That is, we're looking at the concluding end of the search-communication process where a user has communicated a need, and then, must interpret the (usually) overwhelming set of results. Providing genre information might help in filtering the output, and helping with both efficiency and efficacy, as well as satisfaction with the experience.

Towards this end we have designed a set of experiments to test our premise, and have built a corpus of genre-tagged webpages to populate our test collection. The identification of the genres was carried out as follows: We elicited names of genres from respondents from three domains (teaching, journalism, and education) who identified the genres of pages they visited while working on a real task for their own work. We recorded clues and labels, and roughly organized the genre terms into a shallow hierarchical taxonomy so that we could manipulate the granularity if needed.

Next, we used a clustering search engine to harvest potentially useful webpages for a set of 12 "canned" tasks/questions that we think will provide good opportunities for testing the effectiveness of genre information. A clustering search engine (such as clusty.com) uses a proprietary algorithm to present search results in groups that can roughly be understood to be "topical." We say "roughly" because the clustering is in fact only approximately topical, and at times the logic behind it is not immediately apparent. But, overall, we can see that these clusters are keyword based.

Once we harvested the roughly clustered pages, we tagged each one with a genre identifier using the previously devised genre taxonomy (from the field experiments) as well as supplemental terms as they were needed. These terms were chosen by the page analyst. In other words, our labeling decisions for the purpose of this study were not intended to be a general-use genre taxonomy. While we had thought originally that we would be able to do so, in the process of developing a controlled experiment we found we had to find appropriate pages first, and label them second, not always following the genre terminology elicited from our informants. We imposed both labels and structure in order to make the two experimental conditions comparable.

Thus, we aimed to make our genre terms understandable (which we'll test in our pilot), but not necessarily 100% analogous to those we had originally collected in the field experiments.

In the experiments, which we have not yet conducted, these pages will be presented to the subjects in two ways: clustered by keyword (much as they are clustered by the search engine), and clustered by genre, using the terms from the taxonomy as well as additional genre terms as needed.

We have run into many issues, several of them so well articulated in the call for papers for this colloquium:

- It is often difficult for people to identify genres by name, even if they are clearly identifying a "genre."
- Genres differ in their inclusiveness and specificity
- There are many equivalent or near-equivalent terms for the same webpage
- It's difficult to unambiguously link a genre, to a task, or to the clues that were identified.
- Genre-identified webpages can be composed of other genres.
- Some "genres" are highly personal and idiosyncratic.

We report on the practical, working solutions to these questions for the purpose of creating an *experimental* (i.e., controlled) corpus. We have structured this corpus in an admittedly somewhat artificial way so as to provide the maximum control for our experiments. At the same time we have tried to preserve the links to our "naturally" elicited genre terms in order to ground the experimental environment in what people actually do. In doing so, we recognize that much rich genre information was either too difficult to represent or had to be pared away. While our solutions were out of necessity pragmatically driven we feel they may still offer possible guidance for design principles in the future. The building of this corpus was a particularly vexing and complex problem and thus, we would welcome the opportunity to share our experience with others and learn from them.