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

Barbara H. Kwaśnik[1], Kevin Crowston[1], Joseph Rubleske[1] and You-Lee Chun[1]

[1]Syracuse University School of Information Studies (USA)

Short Abstract

This presentation 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. We have built a corpus of genre-tagged web pages and structured this particular experimental corpus in such a way as to provide the maximum control for our experiments. We recognize, however, that much rich genre information was either too difficult to represent or had to be pared away.

Long Abstract

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, thus improving 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. In keeping with the colloquium's theme, our presentation is not about the experiment per se, but rather about building the corpus to be used in the experimental conditions. 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 14 "canned" tasks/questions that we think will provide good opportunities for testing the effectiveness of genre information. An example of a task is to find a webpage that lists the countries that became independent on the same date as Guatemala. A clustering search engine (such as clusty.com) uses a proprietary algorithm to present search results in groups that can

[i] This research was partially supported by Grant 04–14482 from the US National Science Foundation.

roughly be understood to be "topical." We say "roughly" because the clustering is in fact only approximately topical since it is based on keywords and at times the logic behind the choice of the keywords is not immediately apparent.

Once we harvested the roughly clustered pages (about 200 pages per task or 2800 in total), we tagged each page with a genre identifier using the previously devised genre taxonomy (from the field experiments) as well as supplemental terms as they were needed. The genre terms were chosen by the page analyst. A small number of pages could not be coded because the analyst could not identify a genre.

In the experiments, which we have not yet conducted, the pages will be presented to the subjects clustered by keyword (as they were clustered by the search engine) and clustered by genre, and performance on the tasks in the two conditions will be compared (e.g., time, number of categories and pages examined, quality of solution and overall satisfaction).

In developing the corpus of webpages for the experiment, 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 pieces of other genres.
- Some reported "genres" are highly personal and idiosyncratic.

We report on the practical, working solutions to these questions for the purpose of creating our *experimental* (i.e., controlled) corpus. 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.

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.