Stalking the Wild Web Genre (with apologies to Euell Gibbons)  
by Mark A. Rosso

Genres – they are not the offspring of the taxonomist’s notebook or traditional thesauri, but rather, they originate from people’s everyday speech and usage of information. Genre names are useful for referring to communication in the recurring situations in which we find ourselves – whether it’s looking for a job/hiring someone (a resume) or acknowledging a thoughtful gift/having that gift acknowledged (a thank-you letter). How can this everyday terminology be useful for web retrieval?

Imagine being able to tell your web search engine the types of pages you are looking for and even types that you don’t want to see: that ability is the vision of incorporating genre into the interface of web search engines. Genre could be part of the query formulation or a description of each result on the search results page. It’s an idea that’s been around for a while. But we’ve yet to see an implementation of retrieval by genre in any web search engines. Research issues have included the identification of specific web genre labels and genre classification by automated algorithms. For the last decade or so, academicians have been pursuing the idea with mixed results so far. Is it time to give up the pursuit, or is there still uncharted territory out there to discover? Let’s look at the pros and cons of this somewhat controversial research area.

Reasons to Investigate Web Retrieval by Genre  
The following are among the reasons to investigate web retrieval by genre:

- **Genre is used to retrieve physical documents.** People normally store and retrieve documents by genre. We all have places to store our books, old income tax forms, receipts, etc. Certainly, libraries and archives make good use of genre as well. We will return to this point at the end of the article.

- **The web has already spawned unique genres, while many traditional genres have migrated to the web.** While poems, recipes, newspapers and other formats are now on the web in versions still similar to their print lives, totally new web genres have sprouted: homepages, blogs, FAQs and wikis. Information seekers should be able to exploit the obvious differences between these document types in the digital realm.

- **There is a need for non-topical search descriptors.** Keywords that we type into Google are typically representative of topic or subject, and we all know that search needs are more encompassing than that. Although a savvy searcher can often use keywords to indicate the contextual aspects of an information need, a document’s contextual information is generally not easily predicted by the words it contains. It makes intuitive sense that our ability to precisely describe our information needs must increase as the web continues to grow. We can be pretty certain that our patience to sift through 10 or 20 search results will not grow at the same pace.

- **Studies have shown that people can recognize the genre of digital documents.** What’s missing with these new genre are aspects of physical form – they’re all flat on your screen, whereas in the physical world, we distinguish recipes on 3-by-5 index cards from the morning newspaper without any thought. Luckily, research has shown that folks can distinguish certain types of digital genre pretty well. In fact, independent user studies conducted in Sweden, Germany and the United States for the purpose of eliciting the web genres that people conceptualize have identified similar genres [1]. It would seem that shared genre knowledge of web pages has a somewhat cross-cultural basis!

- **Genre is a compact way to describe a document.** With the burgeoning global trend to mobile search from our personal devices, novel ways are needed to present search results on small screens. In just one or two words, a genre label can say a lot about a webpage.
People think this is a good idea. Some of the most popular tags for web pages on the social tagging site delicio.us are genre labels, such as blog, howto, tutorial, news and research. Participants in an experiment that studied the evaluation of genre-labeled web search results said they liked having the genre label there [2]. Lots of researchers, including those who have written in these very pages before, have focused their research agendas on web genre [3]. Several classifications for digital objects, including the Dublin Core, have fields to describe genre-esque attributes such as document type or resource type. FaceTag, a prototype collaborative tagging tool for the web (described in a Bulletin issue last year [4]) uses resource type as one of 10 proposed facets. The Hawaii International Conference on System Sciences (HICSS) has sponsored a mini-track on digital genres for many years now. Finally, the National Science Foundation awarded a team of academic researchers a $150,000 grant in 2005 to explore the use of genre in web retrieval, and there is currently another much larger grant proposal under consideration.

Reasons that Web Retrieval by Genre Might Not Be a Great Idea

The following are among arguments suggesting that web retrieval by genre might not be a workable notion:

- **Not all web pages are the result of typical, recurring situations.** What makes genres so useful is that they embody time-tested communicative actions and reactions to circumstances in which we human beings often find ourselves. However, it makes sense that some web pages are not products of typical situations, or if they are, the searcher may have no idea what that typical situation is. See the next point.

- **Any given web user is not a member of many of the unlimited number of genre user groups represented on the web.** Have you ever clicked on a link on a search results page and gotten something, but you had no idea of what it was? The web gives us access to all sorts of documents that we would never have had the opportunity to view without it. Unfortunately, that also means we can see pages that were never intended for us or that we cannot even remotely understand, lacking the context in which they were developed. If these pages do have genres, they are not ones that have any meaning for us. Their labels may not be of much use to us (or many users) for search purposes.

- **Genre information is already present in many web search results.** Studies have shown that information already present in search results – in the title, the snippet or the URL – oftentimes indicates the genre of the web page. For example, we’ve all seen the tilde in a URL which strongly suggests a personal page. With all those implicit genre cues out there, the benefit of adding a genre label to search results is diluted.

- **Genre is a moving target.** Research in genre theory has shown very clearly that genres evolve over time because the circumstances in which we find ourselves and the acceptable responses to these circumstances change, too. This evolution is especially true today and especially true for the web. Genre research done in 2004 reported on study participants who didn’t know what a blog was. There would be much less of that today in 2008, I’m sure. Why is this genre evolution a problem? It means that the genre labels used to describe web pages and the algorithms used to automatically classify web pages would need to be updated on an ongoing basis. This would be a lot of work for a search engine feature whose benefit is presently unclear.

- **Users may not understand the relationship between genre and their search (or there may not even be a relationship).** During my dissertation research, several subjects made inaccurate relevance judgments based on genre labels in the search results [2]. What they were looking for was on the target web page, but the subjects were of the opinion that a personal page or a blog (as advertised in the search result’s genre label) wouldn’t be the place to find what they were looking for. This skepticism may be a product of the moving target phenomenon noted above. If genres are evolving, it makes sense that people will catch on at different rates, leading some to make wrong decisions based on past assumptions or observations.

- **Even if useful genres can be articulated, it is unclear that machine classifiers are up to the task.** Although many genres have distinctive forms that classifiers could take advantage of, some don’t – they are documents that typically contain paragraphs of text, distinguished primarily by the type or style of verbiage. A news article with a title
and text can look like an executive summary – until you start reading. Given the large number of documents and document types, and the diversity of authors and tasks, web genre may just be too heterogeneous for classifiers to handle at an acceptable level of error.

- **Even if machine classifiers are up to the task, who’s going to annotate all that training data?** Unless we can make a game out of it, like Luis Ahn’s ESP Game, this annotation is an expensive proposition – especially if these moving targets need to be continually maintained! Maybe type social tagging (think del.icio.us) can come to the rescue here…

- **Even if implementation is possible, genres could be spammed easily.** Think personal pages that are just there to sell you life insurance. More seriously, especially if advertising were a genre that one could exclude from search results, many pitches would go underground and masquerade as other genres.

### Conclusion

I can sum up the search for the elusive web genre in three steps. First, appropriate genre labels and definitions must be determined and validated through user studies to ensure that the terminology is indeed understood by users as indicating familiar genres. Second, it needs to be shown that the genres identified are useful for web search tasks. Finally, the genres must be predictable by machine algorithms as there are far too many pages on the web for search engine companies to classify by hand.

Genre is not the only research area that is held back by a lack of usable training data. In fact, some researchers [5] have called for the establishment of a discipline called *annotation science* (www.itl.nist.gov/iaui/894.02/minds.html).

The objective is to develop processes for identifying what needs to be annotated, the annotation process itself (both manual and automatic) and the embedding of best practices into tools to help streamline the process. Genre researchers should definitely take part in this effort.

I want to leave you with a story [6] that was relayed to me by someone who managed an engineering department’s project files for a large corporation in the early 1970s. The company’s capital construction projects, ranging in value from a few million dollars to more than $100 million, on average generated a lateral file drawer for every $2 million of a project. That’s a lot of documentation. Under the original file organization, retrieval success for engineers’ requests for project materials was only about 60%. To make matters worse, it was impossible to tell if the materials needed in the other 40% were misfiled or never even received to begin with. After studying the patterns of engineers’ requests, the manager observed that more than 90% of the time the engineers invoked document type as part of their retrieval request. Consequently, she initiated an experiment to re-organize the project materials by genre and then by company name (equivalent to corporate author) within genre. Retrieval success quickly improved to 96%, and it could be determined with certainty that the other 4% was never received. Organization by genre was the key to success. Ongoing analysis of incoming documents ultimately identified more than 80 unique document types which were applied as the primary organizing structure of all new project files.

The lesson to be learned from this story is that genre can be a powerful hook into the relevance of a document. And, as far as the ever-growing web is concerned, Web searchers may soon need all the hooks they can get.