Chapter 3
Identification of Web Genres by User Warrant

Mark A. Rosso and Stephanie W. Haas

3.1 Introduction

Genre is seen by many as a promising enhancement to the process of web search [4, 12, 16, 23]. The capability to specify or exclude certain types of web pages during a search is intuitively appealing. Historically, document type has proven to be a useful tool for document retrieval (e.g., [6]).

Figure 3.1 graphically depicts how the use of genre in the web search engine interface could enhance web search at two points in the search process: formulation/reformulation of the search query and browsing of the search results.

A genre recognized as relevant to the user's information need could be part of the user's query formulation. For example, a user could specify that only documents of that genre be included in the search results; or, a user might decide to exclude from the search results documents of a genre deemed not to be useful. In either case, document genre is being used to constrain the search space, with the intent of improving the search results. In essence, part of the users' task of filtering search results would be taken on by the system.

The second point at which document description by genre could be helpful is in viewing the search results. Labeling each document description with document genre could help the user to make faster and more accurate relevance judgments, and omit the viewing of some documents' full-text, thus shortening the time needed to assess the documents' relevance. Genre information in the search results could also be useful for query reformulation. For example, a user searching for detailed information on a medical condition, may notice a preponderance of advertisements for products in the search results, and could choose to exclude that genre from future results.

Also, it has been suggested that presentation of search results could be based on the characteristics of the genre of the documents that the results represent: *genre oriented summarization* [8]. For example, the summary of a product review might

M.A. Rosso (✉)
School of Business, North Carolina Central University, Durham, NC 27707, USA
e-mail: mrosso@nccu.edu

A. Mehler et al. (eds.), *Genres on the Web*, Text, Speech and Language Technology 42, DOI 10.1007/978-90-481-9178-9_3,

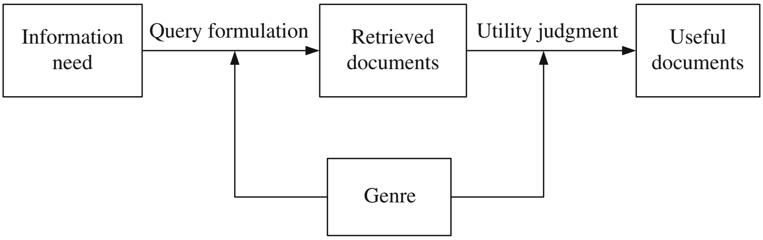

Fig. 3.1 Two points where genre could impact the web search process: query formulation/reformulation and judging search results

mention price and features while a movie review could describe the plot and include the running time of the film.

In addition to these explicit uses of genre to improve the search process, Braslavski (Chapter 9, this volume) suggests an implicit use: incorporating genre automatically into the improvement of relevance ranking of search results.

Thus, genre would seem to be of great use for enhancing search. However, the implementation of information retrieval by genre is problematic on the web – an immense, heterogeneous collection of documents from disparate sources. The pages are generally not labeled with genre metadata, there are far too many for manual classification, and it is unclear how the incredible diversity of the collection will allow for the development of effective automatic classification algorithms.

In addition to these thorny issues, a more fundamental complication exists. Before any classification (manual or otherwise) can take place, the actual genres to be used in the classification must be decided upon. What set of genres can adequately describe to users the contents of the web? What methods can be used to determine (or discover) the members of the set? Given the diversity of pages, the method for reaching this initial decision is far from obvious.

The goal of this chapter is to address methodological considerations in the selection of genre labels to be used to describe web pages indexed by web search engines. For the purposes of this chapter, we will consider that the specification of a genre includes a description or definition (intensional, extensional, or hybrid) of the documents that fall into the genre category, and a name or label that users can recognize as identifying the genre. We first propose criteria for the identification of web genres, and the types of methodologies that are implied by those criteria. We then discuss in detail how the concept of genre applies to the web, and identify the resulting implications for the development of web retrieval by genre. A series of user studies designed to create a genre "palette" are used as examples to illustrate the issues involved in developing a methodology for the identification of genres for enhancing web search, based on the concept of user warrant.

A fundamental difficulty in designing genre studies is how to incorporate the context of search in a realistic manner. Aspects of context come into play at multiple points. The user's context of search includes what has already been seen, both prior to the search and in reviewing retrieved pages. Pages that seem similar to ones that have already been dismissed as not useful may be examined only briefly, if at all.

Pages of a genre type that the user has found useful in the past may be of more interest, at least initially; the opposite is also true. Web pages do not exist in isolation; the pages that link to them and that they link to form a context. A page may make little sense without seeing its predecessor or parent page. So the user's willingness to explore the surrounding pages could affect his judgment of the target page. To simplify the task of web genre identification, we consider here that a label applies to a genre instance of a single web page, as opposed to applying to a website, or other multi-page instantiation. This is consistent with the reality that current search engines deliver search results as individual pages, but we acknowledge that it is an artificial constraint.

3.2 Criteria for the Identification of Web Genre

We propose three criteria for the identification of genres to be used in web page retrieval. First, the users of the system (or some portion of them) must possess sufficient knowledge of the genre to have some understanding or expectations of what it is. Users unfamiliar with a genre will receive no benefit from encountering its label in the search process. The genre must be recognizable to the searcher. Second, searchers must be able to relate the genre to their information needs or tasks, that is, be able to predict if it is likely (or unlikely) to contain useful information. Otherwise, the genre label will not be meaningful to them in the context of their search. Third, the genres must be predictable by a machine-applied algorithm. Because of the size and rate of growth of the web, automatic categorization with a reasonable level of accuracy is crucial. So, in summary, genres for web retrieval must be recognizable by searchers, useful for searchers' information needs, and predictable by machine.

These criteria for the identification of web genres suggest types of methodologies that web genre researchers could employ in their work. For example, to identify typical genres of a specific user group, one might sample members of the group and ask about their typical web usage and what genres come to mind. To show that a specific genre is recognizable, one might ask members of that genre's user group to name, describe or define the genre of specific web page instances in order to see if they agree – thus showing that the user group does possess the shared knowledge that genre theory normally espouses. Thus, measures of participant agreement are used to estimate the strength (in terms of recognizability) of a genre. Any characteristics of form mentioned by the users could also be noted for use as potential features in the development of an automatic classifier.

For insight into genres' usefulness, the users could also be asked to talk about their search needs and what genres they search for. Ideally, to demonstrate the usefulness of the retrieval by genre concept, one would want to directly compare search systems with and without genre augmentation. Then, any number of search evaluation criteria could be used to show the difference between the two. However, the methodology is problematic in that it would require that the automatic genre classifiers to already be developed. (See Chapter 8 by Stein et al. elsewhere in this book for work in this area.)

Alternatives to compensate for the lack of genre classifiers would necessarily detract from the ability to generalize the studies' results, but could still provide useful experimental data. For example, one could pre-label a limited corpus and restrict the queries used by participants in their search sessions. Also, one might observe and record users' search sessions to uncover relationships between searchers' judgments of search result or document relevance, and the corresponding document genre.

Regardless of the specific methodologies chosen, the criteria of recognizability, usefulness, and machine predictability are necessary for the successful integration of genre into the web search engine.

3.3 Operationalizing Traditional Genre Theory for the World Wide Web

We consider a genre on the web to be a pragmatic type (with corresponding form and substance), that is recognized by the genre's "user group", those with a common or shared knowledge of the genre [19]. The genre classification scheme is derived from user instincts, experiences, and preferences, not any given theoretical framework, much like folksonomies and other user-derived or -generated schemes. These contrast with expert-imposed classifications, like the difference between a zoological taxonomy and a lay distinction between pets and wild animals (see Chapter 7 by Sharoff, this volume, for an example of an expert-imposed classification). Regardless of the derivation of the classification, it is necessary to validate the definitions, labels, and application of definitions to web pages by members of the target group, in order to ensure a reasonable level of recognition and agreement. Without the recognition and agreement by the user group, a page type (i.e., a proposed genre) is not necessarily a genre.

The preceding paragraph encapsulates the challenge of transforming the theoretical construct of web genre into an operational definition (as embodied in labels and definitions of web genres) that could be used by content authors/designers, classifiers (human or automatic), and end users of a variety of applications such as information retrieval or content management. We briefly discuss the issues associated with the transformation in order to provide context for the decisions made in the studies described later in this chapter, and the implications these decisions have for the experimental results.

3.3.1 A Genre's User Group

Traditional genre theory almost always includes the notion of a "user group" whose members share some knowledge about the genre, and thus have expectations about its intended use, form, and substance (e.g. [14, 22]). User groups may vary in cohesiveness or restrictiveness of membership criteria. For example, the primary user group of the letter of recommendation for an applicant to graduate school contains writers and readers of such letters, typically faculty members at colleges and universities. They share knowledge of the purpose of the letter, and what the reader expects in terms of its content, formality, and even legal status. The genre may have somewhat limited circulation, and those not in the primary user group are less likely to encounter it or need to use it. If they do, they may use it for uses other than its intended function. For example, a new faculty member may use such a letter as a guide for writing for his/her first letter of recommendation for a student, or a biographer may glean information about a person's life from it.

In contrast, the user group of the newspaper editorial is varied, with the most salient shared characteristic being that they are readers of newspapers, and are likely to understand the difference between an editorial and a news article (although not necessarily). Level of education (beyond some level of literacy), vocation, and other characteristics are not part of the "membership criteria".

It is important to recognize that any one individual is a member of multiple user groups, both broad and specific, and can view a single page from the multiple vantage points the groups afford. Although the purpose of a search is likely to derive from a user's membership in one group, he/she can switch hats rapidly if something of interest to his/her role in another group appears (serendipity).

When we focus specifically on web genres, this view of a genre's user group does not change substantially. There are still cohesive user groups who work with specialized web pages, and have clear expectations of what they contain; these expectations are not widely shared outside the group. Indeed, the web may provide the means for even more specialized groups to exist: profession-based groups, hobby and fan groups, employees of a single company, and so on. They may recognize more specific genres, or have more accurate expectations as to their form and content, even though they are not the only web users to encounter it. For example, anyone can find a university department's home page on the web, but a faculty member at a university may have stronger expectations of what information should (and shouldn't) be there, and how it should be organized than, for example, a high school sophomore. Thus, the "web user", like the non-web "newspaper reader" will have shared experience and expectations about genre-related characteristics of commonly encountered types of web pages.

However, searching the web greatly increases the likelihood that someone from outside of a genre's primary user group will encounter an instance of that genre. Pages from relatively esoteric user groups may turn up in search results, or someone may deliberately seek information that is outside their usual information environment, e.g., a consumer searching for expert health information. Thus, although a page may be created by and for a specific user group as an instance of a familiar genre, the page may be viewed by "outsiders" to whom the genre is foreign. This is a characteristic of web search that genre augmentation may not improve. Another such characteristic of the web is the existence of pages that are not the results of recurring situations, i.e., not recognized by any user groups as belonging to any genre. Ideally, such pages would remain unlabeled by an automatic genre classifier. Complete coverage of a collection (suggested elsewhere in Chapter 4 by Crowston et al., this book) is neither possible nor desirable.

An issue that is related to the concept of user group is the level of abstraction of a genre. Broader genres, such as article and home page, are hypothesized to typically be recognized by larger or more diverse user groups than narrower genres like an SEC filing or copyright transfer agreement [24]. A question for further research is whether the characteristics of the "typical" web user tend to be associated with broad, large-grained genres. In other words, how specific are the genres recognized by most everyone, and how strong are the expectations about the genres? Is the concept of "web user as user group" useful for our purposes? How useful are the broad genres that they may recognize for improving web search? Some researchers have questioned the utility of considering web users as a whole to be a relevant group in terms of retrieval by genre [15, 16].

Despite the issues of pages with unrecognized or unknown genres, one can conclude that the concept of a genre's user group is very much applicable to web genres, and not materially different from documents in other media. What are the implications of this for web genre research? Operationalizing the "user group" as a group of people with obvious shared characteristics, such as profession or workplace, thus also characterizes to some extent the websites they frequent, specifically those associated with the shared characteristics. This is likely to make some parts of the research easier. The limitation provides some justification for limiting the sample of web pages used in the research by domain or organization. The participants are likely to have more shared knowledge (e.g., what an academic department does), familiarity with the work and work documents, and thus be able to recognize more specific genres and have more accurate expectations as to their intended use, form, and content. Because of these expectations, they may also be able to see the utility of using genre as part of information seeking.

However, generalizing research findings to other user groups, or to web users as a whole, will be problematic. Some groups may work with more specific genre that support stronger expectations than others. Some specific genres may have characteristics that are more easily usable by people outside the primary user group for some purposes, e.g., finding links to relevant information.

3.3.2 Genre: Function, Form and Substance

In discussing the individual aspects of the genre pragmatic type and how they apply to the web environment, we set up a sense of distinctness among them that does not exist in reality. In use, the distinction between function, form, and substance blur: form shapes substance, substance entails function, and so on.

Function. The "function" of a web genre could be viewed from two perspectives: that envisioned or intended by the creator of an instance of the genre, and that perceived or acted upon by the user. For a genre used by members of its intended user group, the two perspectives will generally be aligned. Non-members' actual uses of the page may be in alignment, or be entirely different. The common phenomenon of using a genre as a container of needed information, rather than for its intended purpose, frequently occurs on the web. In some of our studies, participants would commonly judge a web page as useful not because of its content, but because it contained a link to the desired content. In this scenario, recognition of the utility of a genre means recognizing evidence of where it might lead, overlaying a directory or referral-type function on top of its intended function (what Chapter 4 by Crowston et al., elsewhere in this book, describe as "borrowed purpose").

Adding to the difficulty is the fact that search engines return individual pages, isolated from related pages that may provide needed context – potentially important pages whose existence may not even be known to the searcher. The function or purpose of a genre is traditionally seen as a shared understanding among creators and users of the genre as to its role in actions and communications. The shared understanding is based on knowledge of the context in which it is used. On the web, the originally intended context of pages can be more elusive: users may come to a page deep in a website from a Google search, and may have little interest in looking beyond it. Any guess as to the purpose of the page is based on face evidence, not an understanding of its context. This type of situation could increase the difficulty for even a genre "insider" to recognize a page's genre. Thus, single-page genre validation methodologies (such as the one described later) could underestimate users' recognition of a genre.

For research into the use of web genre for information retrieval, these observations suggest that asking subjects to rate the utility (or relevance or whatever construct is used) of a genre instance could be misleading. A page could be judged useful because the user views it as supporting a function that is unrelated to the definitional functions of its genre. This does not mean that the user hasn't recognized its genre, or has no expectations associated with it, rather that the user associates different (or additional) functions with it. For example, someone looking for the title of an article written by a faculty member may judge a department home page to be useful, because from there, he can find the faculty member's personal page, which is likely to link to his CV, which should have the article listed. Asking subjects to articulate the reasons for their judgments is more likely to reveal their view of the functions supported by the web page.

Substance. By "substance" we mean the content (which may include topic) of a genre. When experiment participants are asked to name a web document's genre, they often conflate topic and genre. Theoretically, genre labels should be as topic-neutral as possible. In practice, some genres are more closely tied to topic than others. For example, the substance of a newspaper article is a description of an event or situation, usually including information about the people and places involved, and often carrying an aspect of timeliness. Within this substance, however, the range of topics is vast; elections, war, weather, tennis, fashion, or just about anything else. Substance and topic are relatively independent. In contrast, the genre of university course listing is inherently about courses. The substance includes course numbers and titles, and often a brief description. The topic could be broader or narrower, for example, listing only chemistry or sociology courses, but the distinction between the topic and substance is fuzzy.

The substance may be communicated by a series of moves (e.g. [2, 22], or types of information that are typically included in a genre instance. In a letter of recommendation, expected moves include a greeting, a description of how the writer knows the applicant, and reasons why the writer recommends the applicant. The kinds of reasons cited, or how they are framed may differ according to the type of application. For a job, the letter may discuss past educational accomplishments, while a recommendation for an award may discuss why the applicant is worthy. Consideration of the substance of web genre follows the same pattern as the previous consideration of function: a user may use elements in unexpected ways.

A fundamental difference among web genres, traditional genres implemented on the web, and non-web genres, is the presence of hyperlinks. This expands the notion of substance: the link text itself can be substance, but is also a reference to another page. The target page forms some part of the context of the initial page; users may consider its substance to be a part of the initial page's substance. The implications for experimentation are similar to those for function: a user may consider a genre instance useful because of its links and the pages it links to, rather than the page itself. Hyperlinks are also responsible for the research decisions over what constitutes a genre instance: an individual web page, an entire website, or a multi-page document (e.g., an FAQ (frequently asked questions) that spans multiple web pages).

The URL is another web-specific element: users may pick up clues as to the genre of a web page, and therefore trigger expectations of its utility, by words or abbreviations contained there, e.g., "home", "interview", or "syll".

Form. Form is the most obvious difference between traditional and web genres. Web genres do not provide the same physical cues (weight, size, material, etc.) as their traditional counterparts. Nonetheless, research has shown that people can recognize genre [23] and elements of specific genre [5] in digital environments.

Form is the vehicle through which genre function and substance are expressed. Returning to the letter of recommendation, the expected moves may be expressed in casual, formal, or extremely formal language. Form includes whether a letter is typed or handwritten, and even the kind of paper used. On the web, means of expression are practically unlimited, including sound and images, color, escape from the normal (for western languages) top-down, left-to-right scanning, and even form that changes as the user watches. The form of a web page can be indicative of the context of the page: the home page of a university department will use different design elements than a children's game website, although both may embody the directory genre. As page design conventions have coalesced over the past decade, the web user can expect some common elements on most, though not all, pages. The form of some genres and genre instances may be exactly the same as their non-web counterparts, as is often the case with a .pdf document. Other specifically-web genres, such as the home page and the blog, have developed their own conventions. The appearance of the substance elements is as informative of genre as the actual words or pictures themselves. For example, a list of questions at the top of the page that are links is highly suggestive of a FAQ, as opposed to an interview, which typically has alternating questions and answers.

3.3.3 Genres on the Web: Further Implications for Research

In many respects, traditional genre theory transfers easily to the web environment. The aspects of function, substance, and form are still integral to the definition and expression of a genre. The user group is also essential to the core definition of a genre, but the digital, accessible, and linked nature of web documents provides more opportunities for people outside of a genre's primary user group to view instances of the genre. These considerations affect both the selection of experiment participants, and the construction of the sample of web pages.

The presence of hyperlinks is the other important distinction between traditional and web genres, which impacts research design decisions. The perception during a web search that a document may link to something useful may have nothing to do with the document's genre: it's simply functionality added to a document (and not its genre) by the existence of hyperlinks. In other cases, the linking expands the context in which a page is viewed. For example, organization home pages can link to individual person's home pages.

These distinctions suggest that genre researchers who observe users' web search behavior, must gather more information about page utility from users than just a bare rating. The reasons for the judgment may reveal that the page itself isn't useful except as a starting point: the links to related pages (i.e., the page's context), and expectations about the related pages may be the reason for a "useful" rating. Further, a genre instance may be implemented on the web to span multiple pages. For example, a frequently-asked questions page (FAQ) may have the questions on one page, and answers on separate pages, yet users may perceive them as a single "document". Researchers must decide if subjects should be allowed to follow links when making utility judgments, and if so, how far afield they may go.

3.4 Developing a Web Genre Palette

As web genres are recognized by their respective user groups, the collection of terminology to describe web genres would, ideally, directly involve the users. At a minimum, proposed genre terminology (labels and descriptions) would be validated by users in order to show that the identified labels do indeed represent genre. Thus, the genres are identified by user warrant, meaning that the appropriateness of the terminology is affirmed by the users' actual use of the terms.

A series of three user studies [19] was undertaken with the purpose of developing a genre palette for use in web retrieval. In order to start on a more manageable problem, pages to be examined by participants were limited to the edu domain, as in Rehm [16]. The web pages in the terminology studies were collected by interval sampling the Google search results obtained from one-word queries consisting of the most frequently used English words [7]. As discussed earlier, the choice to restrict the user group not only limited the pages that could be included in the sample, but also limited the generalizability of the results. The choice was made partly to avoid problems that earlier studies attributed to a web-wide focus: that it leads

Table 3.1 Overview of the studies

	Methodology	Product
Study #1 Survey of user terminology	3 participants individually separated 100 webpage printouts into stacks according to genre, assigning names and definitions to each genre	A collection of 48 genres names with definitions
Study #2 User-based refinement of terminology into a tentative genre palette	10 participants individually classified 100 webpages (same as in the previous study) using the 48 genres (plus a "suggest your own") category	A palette of 18 genre names and definitions
Study #3 User validation of the genre palette	In an online experiment, 257 participants each classified a new set of 55 webpages using the 18-genre palette	Validation of participants' ability to classify pages using the palette
Study #4 Measurement of user relevance judgments of genre annotated search results	32 participants performed 4 tasks. In each task, participants judged the usefulness of 20 search results and 20 web pages according to an assigned task scenario	Comparison of participants' performance with and without genre annotated search results

to vague and unusable results. We also desired to minimize the size of the resulting genre palette so that if the palette were used in search engine query formulation, the choice of genres available to the user in the search interface would be a manageable number. Finally, a fourth user study was conducted to gauge the usefulness of the genres identified in the first three user studies for the purpose of web retrieval. See Table 3.1 for an overview of the four studies.

The intended user group, people who share genre knowledge of web pages in the edu domain, was operationalized as college graduates. Arguably, a college graduate is most likely not as aware of the workings of an academic department as a departmental staff member would be. It is recognized that this experimental design choice, obviously made for convenience, could impact the validity of the results.

3.4.1 Collecting Genre Terminology in the Users' Own Words

In the first study, three participants (an information technology professional, an organ transplant social worker and a computer science professor), in separate sessions, were given a stack of 102 web page printouts, and were asked to separate the pages into piles according to genre. They were also asked to name the genres by writing the names on sticky notes and placing them on the piles. After the piles were complete, participants were asked to provide a short, one or two sentence, description of each genre, and then to describe the page characteristics that led them to place a page in that genre. Participants were also asked to identify the most and least representative pages in each pile, and to explain those choices. At any time during their explanations, they were allowed to move pages between piles, and to explain these moves.

Major experimental design decisions made here include how to present the pages to subjects, and how to allow them to name and group the pages. Certainly, allowing participants to interact with the pages in a web browser would establish a more realistic context for their experience of the pages. In addition to the fact that perusing 8.5″ by 11″ pieces of paper is not the natural way to view web pages, other compromises had to be made as a result of the printing. Page backgrounds were not printed because that inhibited the readability of many pages, as well as using a lot of ink. Web pages consisting of multiple printed pages were stapled together in the upper left margin of the printed pages. Long web pages (i.e., in excess of 10 printed pages) had middle pages (mostly with repetitious content and/or formatting) excluded from the printing. As genre is characterized by specific types of content and format, we hypothesized that these omitted pages should not have materially impacted the subjects' assessments. Some pages were omitted from the final sample for various reasons. Some pages looked radically different in print (often because of the missing background). Some pages just would not print properly. Despite the use of color printing, in some cases, it was hard to discern what text represented links. Thus, it is possible that participants' terminology did not fully take into account the importance of hyperlinks noted earlier in this chapter.

Despite the obvious limitations of using printed web pages, the printouts provided the participants with tangible things to place in piles (which they could name, give definitions to, and move pages between, easily and whenever desired). We did not have the resources to construct a software-based alternative that could have provided this much functionality for implementing a "card-sorting" process (e.g. [17]) with web pages, and it could be argued that users unfamiliar with the software would not find the online "piles" as hospitable to rearrangement as physical piles.

The session lengths ranged from 1.75 to 2.5 h, and still some genre names, definitions, and sorting decisions were left unexplored. It is our perspective that this was an effective, albeit time-consuming, method for gathering the desired genre terminology. Thus, we made the design decision to limit the sample size of this first study to three participants.

A danger of using such a limited participant sample is overfitting the results to this specific sample. Our experimental design reduces this possibility by filtering the resulting genres through the two subsequent studies. In the second study, a new participant sample gives their input on the genres named in the first study, and a refined set of genres is created. This refined set of genres is then given to a third set of participants for describing an entirely new set of webpages.

In this first study, the three participants used similar wording or concepts for their piles' names and descriptions, in many cases. For some pages, participants grouped them at different levels of abstraction (e.g., one had separate piles for FAQ and Help, while another had a combined FAQ/Help pile). In addition to the genre names and definitions collected, the page characteristics (in [19]) that participants associated with specific genre could be helpful in building automatic genre classifiers.

Note that the card-sorting process does not allow participants to place web pages in more than one pile. For example, if a home page contained a search box, the participant was forced to choose between the two genres, home page and search engine. This is clearly not a realistic categorization. Many researchers have noted that web pages can contain elements of multiple genres (e.g. [9]). However, given that the purpose of this first study was to collect genre terminology (i.e., names and definitions), the particular categorization of any given page was of secondary importance. We do acknowledge that this restriction could have affected the names and definitions that the participants generated.

The principle of user warrant requires a generation stage in the development of a genre palette. The card-sorting technique clearly demands a lot of effort from the participants, but the method used here allowed them to find similarities among pages first, and then name them. Thus, their genre definitions were based on several instances of what they viewed as a genre. In contrast, the method used in Chapter 4 by Crowston et al. (elsewhere in this book) asked users to generate a genre name as they viewed each individual page, which may be a more difficult task. Either way, the generation stage must be followed by a refinement stage, to group and normalize genre names.

Genres names elicited from the participants included familiar document types such as article, abstract, bibliography, course description, job listing, newsletter, etc. We crafted the terminology from this study's three participants into a list of 48 genre names and definitions, keeping the terminology as similar as possible to the original, while combining definitions which were nearly identical in wording. Many of the genres left in the list were still quite similar (e.g., product for sale, and shopping). The rationale for this is that genres, if expressed in user-generated terminology, should theoretically be more easily recognized by members of the genres' user group. For the complete list of the 48 genres, (see [19]).

Given the frequently synonymous and overlapping definitions in the list resulting from this study, the goal of the next study was to help refine the terminology into a smaller set of mutually exclusive genres.

3.4.2 Users Choose the Best of the Collected Genre Terminology

In this second user study, the extent of user agreement would once again be used to determine the most natural terminology, but this time with a different set of users who would vote on the terminology collected in the first study. Each of ten participants was given the list of genre name/definition pairs, the same stack of 102 printed web pages (arranged in a different random order for each participant), and a data collection form to record a genre for each web page. For each of the 102 web pages, the participant wrote a number from the list corresponding to a genre/definition pair which best described the page; or suggested his/her own genre name and definition, if none of those in the list seemed adequate.

The participants were drawn from a convenience sample of approximately 10 college graduates of various occupations. The ten sessions ranged from 65 to 120 min, for an average of 90 min per session overall. From a list of 49 genres (including the addition of the "none of the above" option), many of which were extremely similar in nature, the resulting level of agreement is quite acceptable: half or more of the participants agreed on one genre for a given page in 60% of the instances. This result is particularly notable, given that each of the 10 participants was voting on terminology from three other people, all collected independently from each other.

Another factor that might be detrimental to the agreement level here is that the definitions shown to the participants in this second study were presented out of context. In the previous study, each genre definition was part of a participant's constructed genre palette. If we think of a palette as a collection of genres, each genre definition not only describes a single genre but also impacts the boundaries of other genres in the palette. That quality was lost in this study in which several palettes had been combined. Unlike a genre definition in a genre palette, each definition in this study had to stand on its own. These genre definitions can also be considered to be out of context because the participants in the previous study did not necessarily intend for their definitions to be understood by a public audience.

Of course, as in the first study, web pages presented individually are automatically out of context, devoid of the links to other pages, and pages that link to them. The fact that shared genre knowledge is based on understanding the context in which it is used, makes the level of agreement on genres here seem even more robust.

Another limitation in these studies is the use of the same set of 102 pages in the first two studies. This could work to reduce the generalizability of the resulting palette to other sets of pages. The decision to use the same set of pages again was based on convenience, and may have worked to increase the level of agreement observed.

After the 10 participant sessions were completed, we then developed a set of five principles [19] for creating a genre palette from individuals' sortings. Based on those principles, the original list was trimmed down to 18 genres (see Table 3.2).

Note that the genres in Table 3.2 seem to be at varying levels of abstraction. There are broad genres such as Article and Welcome/Homepage, and more specific genres like job listing and course description. Certainly, the genres named by participants were influenced to some extent by the specific pages in the 102 page sample. Regardless, genres' varying level of abstraction raises research questions for each of the three proposed criteria for genres to be used in search.

First, as noted earlier, what are the levels of abstraction of genres that the "typical" web user recognizes? Does targeting all web users for the user group (i.e., the "lowest common denominator") limit the palette to broad genres? It is obvious that targeting a narrower user group (e.g., people familiar with higher education) does not limit the palette to sub-genres. They recognize all the broad genres that the larger group understands (like article), and even more specific ones like "job listing" that are not specific to the edu domain.

Second, is there a general relationship between genres' level of abstraction and their usefulness for searching? For example, the concept of product review has

Table 3.2 Palette of 18 genres

Genre	Description
Article	Something about a topic, often with supporting facts or opinions
Course description	What's covered in a course; syllabus
Course list	Page that lists courses
Diary, weblog or blog	A personal narrative or time log of activities (not a biographical article)
FAQ/Help	Frequently asked questions, or assistance in helping you perform a task; questions may be links to answers, or topics may be links to assistance; not interactive like a forum
Form	Page primarily for entering and submitting information (other than a search engine)
Forum/interactive discussion archive	One or more messages and/or responses that are viewable by an audience
Index/table of contents/links	A page which is primarily a list of links or text items ordered (usually alphabetically) so that a list item can be found easily, AND the page does not belong to any of the other categories
Job listing	Describes one or more jobs that are available
Other instructional materials	Materials (other than a syllabus) used in teaching courses, including but not limited to tests, quizzes, assignments, answer keys, etc.
Personal website	Page (possibly a home page) that somebody writes about oneself (but not a biographical article)
Picture/photo	Page primarily containing a picture or pictures with few or no words (other than captions)
Poetry	Contains poetry or similar wordplay
Product for sale/shopping	For purchasing products (not a product review article)
Search start	Page primarily to enter key words and search a database; a search engine
Speech	Text of a speech
Welcome/homepage	Starting page (does not have to be the "top" page in a site); may contain introductory information about a specific organization, department, program, etc. and a table of contents
NONE OF THE ABOVE	Page that definitely does not fit into any of the above categories

more distinguishing characteristics than that of article. Does that mean that users could more easily relate product review to their information needs than article? Certainly, it depends on the task and the document collection. In general, though, it makes intuitive sense that broader genres may not be as useful for searching as those sub-genres with greater number of distinguishing features. Lee [13] provides an in-depth discussion about genres' level of abstraction. In a project to label the British National Corpus (BNC), Lee asserted that the level of abstraction does not matter as long as the categories are found to be useful. However, his statements were made in the context of researchers selecting texts from the BNC for linguistic study. For our purposes, this remains an open research question.

The article genre is an interesting case in point. The name can refer to wide variety of documents, from a research article to a newspaper article. One can further subdivide these, recognizing distinctions between a hard news article and a fashion article, or a biochemistry research article and a literary theory research article. The interplay with the user group suggests that multiple levels of specificity might be useful. If a user is in his/her role as general web user, then the ambiguity of "article" may be helpful in making a broad distinction between an article and a FAQ or job listing. The finer distinctions between different subgenres of research articles are not likely to be meaningful to the general web user, whereas they may be important to a researcher. The researcher user group can recognize the characteristics of a typical biochemistry research article. If both broad and narrow genres are useful at different points to different user groups, this suggests that a palette with hierarchical structure would be more adaptable.

Finally, how does a palette containing genres of varying levels of abstraction affect the ability of automatic classifiers? Some researchers have suggested this to be a problem, (e.g. [20]). It makes intuitive sense that a mix of broad and narrow genres could cause problems for automatic classification.

We will re-visit the issue of varying levels of abstraction of the genre palette derived from user terminology. For now, we will turn to the third study. Its' objective is to validate the palette by measuring the agreement among a new set of participants using the palette to label a completely different set of web pages.

3.4.3 User Validation of the Genre Palette

The first proposed criterion for genres to be used in search is that of recognizability by the community of persons (the user group) that create and use the genre in the context of a recurring situation. We operationalized recognizability in this study as the level of agreement between participants in classifying a set of web pages into the genre palette. Agreement is measured on a page-by-page basis by a simple percentage of all the participants' votes. We based this decision on the principle of user warrant. Historically, user warrant was used as the justification for including a term in an indexing system because the users used it to search for documents (e.g. [1]). Although the genre names were not derived from actual searches, it was derived from users' classification activities, which is essentially what people do when specifying a search query: produce terms that describe a document. Thus, we believe this analogy is appropriate. The next decision was to define a threshold of agreement that would represent a sufficient level of recognition. We propose that if 50% or more of the participants say that a page is an instance of a specific genre, then it is. The rationale behind choosing 50% is that it guarantees that the genre is the most frequently cited for that page. In most cases, the genre garnering the second highest level of agreement had much lower agreement than the highest one. We would consider our genre palette as a whole to be "validated", thus satisfying the first proposed criterion for web genres, if the majority of the pages reached or exceeded the 50% threshold. At a minimum, we hoped that the palette contained at least some genres that met the threshold in order to "certify" them as true genres.

A new set of 55 web pages was collected using a method similar to that for collecting the 102 pages used in the first two studies. We created a website to collect demographic data and participants' genre choices for the 55 web pages. After completing the study, participants had the option of giving feedback about their classification experience and/or leaving contact information if they wanted to talk about their experience.

Again, the intended user group was people familiar with the higher education environment. This time, it was operationalized as faculty, staff and students at 4-year institutions. Two hundred fifty-seven people participated in the study.

A flaw in the experimental design was in not collecting enough demographic information regarding the academic disciplines that the participants were associated with. We were not able to determine if the results from this self-selected sample were from a representative cross-section of the intended user group, or biased toward those who may be especially interested in web pages, e.g., people in information technology and information science-related fields.

In any case, the results were quite good. Eighty-seven percent (48 of 55) of the pages reached the 50% recognizability threshold. The average agreement for the most frequently genre assigned for a page was 71.9% for all 55 pages. Inter-participant agreement was 58.3%, with a Cohen's kappa of 0.55. We used two measures to estimate the strength of the individual genres' recognizability.

First, for each genre, we looked at the average agreement for that genre over the pages that were determined, according to our threshold, to be of that genre. The higher this percentage, the more frequently a page of this genre was recognized as being a page of this genre.

The second measure can be thought of as a measurement of "false hits". This was the percentage of votes for a particular genre, across the subset of 48 pages in which this genre was not the threshold-exceeding genre. (Remember that only 48 of the 55 pages received votes exceeding the 50% threshold for any single genre.) The lower this percentage, the less frequently a particular genre was confused with the other genres. In other words, this measure shows how well participants recognized that pages were NOT of this particular genre. Note that this measure of recognizability is imprecise in that all false hits are not created equal: confusion between two similar genres like syllabus and course description is not as severe as confusion between two more dis-similar genres like poetry and job listing

For an example of how the two measures were used, the genre *job listing* scored high in recognizability on both measures: average participant agreement on job listing pages was 82.1%, while false hits were just 0.0%. Together, these two measures gave a more complete picture of the strength of the genres in the palette. An open question is how to combine these measures into one measure. Using these two separate measures, it is not possible to rank these genres according to the single construct of recognizability. Some genres had high levels of agreement, but also more false hits, and vice versa. For example, "course description" had the highest consensus of all the genres at 94.2%. However, it had one of the worst false hit rates at 2.4%. See Rosso [19] for additional details.

Using the two separate measures, we attempted to derive general ranges of recognizability for the genres in our palette. Highly recognized genres included picture/photo, job listing, poetry, product for sale/shopping, FAQ/Help, "diary, weblog, or blog," and search start. Personal web site, forum/interactive discussion archive, and form fell into the medium range of recognizability. Genres with low recognizability were article, index/table of contents/links, other instructional materials, and none of the above. Genres with disparate scores on the two measures were course description, course list, welcome/homepage and speech. These are harder to place in a range, but course description and course list would likely fall into the high or medium range, and welcome/homepage and speech into the medium or low range.

What jumps out from this list of rankings is that the broadest genres (e.g., article) received the lowest recognizability scores. If this finding is corroborated in future research, it has important implications for the future direction of research in web retrieval by genre. We have already said that the usefulness of broad genres for retrieval is an open question. If typical web users are not clear on these broad genres (i.e., there is not strong shared understanding), then it seems more unlikely that they will be useful for search. If that is the case, are there enough narrower genres recognized by the typical web user to make web search by genre feasible? It is possible: in this study, most of the better-recognized genres are narrow, but not specific to the educational domain.

In addition to participant agreement, an abundance of detailed "de-briefing" comments written by participants provided a rich lens through which to interpret the results. Some comments noted the general ease of the task, but participants also noted several difficulties that have implications for the design of future studies.

Some pages fit into more than one category, for example, a home page with a search engine on it. As mentioned earlier, the operational decision to force participants into a one genre per page classification simplified the calculation of participant agreement. However, it made the task less natural. Agreement might be higher if multiple genre assignment was allowed, but it is unclear how that agreement should be calculated.

Another problem noted was that some pages didn't seem to fit any of the categories. Participants suggested many names for these types of pages. This could be an artifact of using a different sample of web pages in this last study. Studies similar to the first two studies may need to be repeated to capture as many of the commonly used genres as possible. Participant comments also suggested that several of the broad genres such as article, other instructional materials, and form should be broken down into more specific categories.

Also, some labels for web page types may not represent a single shared understanding – in other words, the label means different things to different groups of people. For example, one participant made the following comment:

> I found the welcome/homepage a bit disconcerting. Many pages seemed that they were welcome pages, but definitely not homepages, wheras [sic] others were in fact homepages. [18, p.115]

In an email exchange with this participant, it became clear that he considered a welcome page to be a top-level entry point to a website, and a home page to be a personal Web site. A search of home page definitions using Google uncovered both definitions for home page. (Perhaps Dillon and Gushrowski's [5] "personal home page" would work better in the palette than personal Web site.) The point is that some commonly used labels may appear to be genres but that a single shared meaning for the label has not yet crystallized within the user group. Blog is another label that is commonly used to refer to pages with vastly different functions [11].

In summary, although several genres with high levels of agreement were identified in these studies, further user studies are necessary to collect additional genres and to refine genre names and descriptions already in the palette. Questions remain regarding the identification of broad genres with good recognizability, and the decomposition of broad genres down into narrower ones. Methodological issues such as allowing users to assign pages to multiple genres, and how to measure agreement in these cases, as well as the creation of a single measure of recognizability, also deserve attention.

There is still cause for optimism regarding the genre approach to web search. Interestingly, the genres in this palette, although developed independently, are similar to 7 of 8 Internet-wide genres based on user input reported in Stein and Meyer Zu Eissen [21], and similar to 8 of 11 Internet-wide genres as reported in Karlgren et al., [12]. Based on these observations, one might infer that some substantial amount of genre knowledge exists among users, even from different cultures (in this case, the United States, Germany, and Sweden). See Rosso [19] for a side-by-side comparison of the palettes.

3.4.4 A Fourth Study: Determining the Genres' Usefulness for Web Search

Having identified a palette of fairly recognizable genres in the first three studies, the next step was to investigate whether using genre to augment web search could produce a noticeable improvement. The final study compared participants' ability to make relevance judgments of web page search results with and without the pages' genre label included in each search result. Thirty-two participants (college faculty and staff) performed 4 tasks in random order. In each task, participants judged the usefulness of 20 search results and 20 web pages according to an assigned task scenario. The stability of each judgment from search result to actual (the "gold standard") was measured. Search results were labeled with the genre of the web page in two of each participant's four tasks.

Overall, genre-annotated search results did not produce faster or more stable relevance judgments. However, many users preferred having the genre of the web page available in the search result to help them in the evaluation process [18].

What do these results mean for this line of research? There are many possible reasons for not finding a measurable difference in performance between genre-annotated search results and "standard" ones. Certainly, tasks, users, collections, and their interactions are all complex variables. In these experiments, the user tasks were assigned, and they weren't real search tasks – each task was a series of judgments of single surrogates followed by a series of judgments of web pages. The set of tasks was long – an average of 1.75 h. Also, participants were not informed that genre labels would be present in half of their tasks; over half of the participants reported that they didn't remember seeing any of these labels!

But comments from two participants of the study described in Section 3.4.3 may yield some insight into how to improve the design of this type of study.

> The category of a page is hardly a consideration when "Where's the information?" is the purpose of the visit.
>
> Normally, I wouldn't seek to classify web pages in order to know whether they were relevant to my interests or objectives; either the information would interest me or not, continue to inform me or not, and I'd move on to the next search technique. [18, p. 116]

These comments echo our earlier discussion about the function of "web page as a container of information" being overlaid onto the function of a page as expressed by page's genre. This study required participants to make relevance judgments on a scale of 1–4, without taking into account the reasons behind the judgment. Relevance judgments may have nothing to do with a page's genre, and everything to do with the presence of the sought-after information. The point is that the influence of genre on the evaluation process cannot be teased out of the experimental results unless it is determined which judgments were made on the basis of genre (and which were not).

Thus, experiments hoping to measure the effect of genre on the evaluation of search results need to include some method for getting this information from the user, while at the same time minimizing the disruption of the user's decision-making process. Methods could include a think-aloud procedure, or a debriefing immediately following the experimental procedure.

3.5 Conclusion

We have described the issue of identifying genres on the web for the purposes of web retrieval. Through the examination of genre theory and the literature on web genres, we have attempted to document the methodological considerations necessary for this research area to progress. More user studies need to be done to collect appropriate genre names and definitions, and to refine those that have already been collected. The issue of broad versus narrow genres, and their usefulness for search need to be explored. Finally, techniques for accurately predicting the genres identified need to be developed.

Several important questions remain. Is there enough social agreement on web page types for this endeavor to be feasible for a web-wide audience, for the "typical" web user? If not, how could this be implemented for smaller user groups? Certainly, corporate intranets with their more homogeneous sets of users, tasks, and pages would be excellent places to start. However, other than corporate intranet users, is there some subset of users that could benefit from search by genre, and if so, what would that solution look like? Would it involve narrow genres of web pages that just these users understand and use? Or would it involve categorizing only websites of interest to this group?

It is worth noting that in finding a web-wide solution, the pages that would be annotated with genre labels are most likely only a small segment of the web: search engines only return the most popular pages. If the solution is built on top of an existing search engine, then only those pages need be annotated by genre. Does the practice of only returning the most popular pages affect what genres are available through major search engines? Would we see other genres if we could find the less popular pages? This is not to be taken as a criticism of the major search engines. They are in business to help people meet their information needs, not to provide equal opportunity for every web author's pages to be found. But, if academic researchers are to make progress in this, or any area of web search, we may need the help of commercial search engines. Others have expressed this concern:

> The commercialization of web search has caused a significant shift in the balance of knowledge between industry and academia; large web search engines have Web data, user data, and computer hardware that researchers cannot begin to reproduce, raising concerns about the quality and relevance of some areas of academic research [3].

Finally, this research area is not the only one held back by the annotation problem. Well-respected experts [10] have called for the establishment of "Annotation Science" to help solve the widespread need of several disciplines for labeled corpora, including developing methods for determining what the labels are. Researchers in web genre should be part of this effort.

References

1. Anderson, J., and J. Perez-Carballo. 2005. *Information retrieval design*. St. Petersberg, FL: Ometeca Institute.
2. Bhatia, V. 1993. *Analysing genre: Language use in professional settings*. London and New York, NY: Longman.
3. Callan, J., J. Allan, C. Clarke, S. Dumais, D. Evans, M. Sanderson, and C. Zhai. 2007. Meeting of the MINDS: an information retrieval research agenda. *SIGIR Forum* 41:25–34.
4. Crowston, K., and B. Kwasnik. 2003. Can document-genre metadata improve information access to large digital collections? *Library Trends*, 52:345–361.
5. Dillon, A., and B. Gushrowski. 2000. Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of American Society for Information Science* 51:202–205.

6. Fidel, R. 1991. Searchers' selection of search keys: I. The selection routine. *Journal of the American Society Information Science* 42:490–500.
7. Francis, W., and H. Kucera. 1982. *Frequency analysis of English usage*. New York, NY: Houghton Mifflin Co.
8. Goldstein, J., G. Ciany, and J. Carbonell. 2007. Genre identification and goal-focused summarization. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 889–892. New York, NY: ACM Press.
9. Haas, S., and E. Grams. 2000. Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of American Society for Information Science* 51:181–192.
10. Harman, D. 2007. Meeting of the MINDS: Future directions for human language technology executive summary. http://www.itl.nist.gov/iaui/894.02/minds.html
11. Herring, S., L. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 38nd Annual Hawaii International Conference on Systems Sciences*. IEEE Computer Society Press.
12. Karlgren, J., I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. 1998. Iterative information retrieval using fast clustering and usage-specific genres. In *Eighth DELOS workshop – user interface in digital libraries*, 85–92. Stockholm, Sweden, October 21–23, 1998.
13. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through The BNC jungle. *Language, Learning & Technology* 5:37–72.
14. Miller, C. 1984. Genre as social action. *Quarterly Journal of Speech* 70:151–167.
15. Nilan, M., J. Pomerantz, and S. Paling. 2001. Genres from the bottom up: What has the Web brought us? In: *Proceedings of the American Society for Information Science and Technology Annual Meeting*, 330–339. Washington, DC, November 2–8, 2001.
16. Rehm, G. 2002. Towards automatic Web genre identification. In: *Proceedings of the 35th Annual Hawaii International Conference on Systems Sciences*, 1143–1152. Los Alamitos, CA: IEEE Computer Society Press.
17. Rugg, G., and P. McGeorge. 1997. The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 14:80–93.
18. Rosso, M. 2005. Using genre to improve Web search. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, NC. http://ils.unc.edu/∼rossm/Rosso_dissertation.pdf.
19. Rosso, M. 2008. User-based identification of web genres. *Journal of the American Society for Information Science and Technology* 59:1053–1072.
20. Santini, M. 2006. Common criteria for genre classification: Annotation and granularity. In *Proceedings of the Workshop on Text-Based Information Retrieval Held in Conjunction with the European Conference on Artificial Intelligence*.
21. Stein, B., and S. Meyer zu Eissen. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*. Ulm, Germany.
22. Swales, J. 1990. *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.
23. Toms, E., D. Campbell, and R. Blades. 1999. Does genre define the shape of information: The role of form and function in user interaction with digital documents. In: *Proceedings of the 62th American Society for Information Science Annual Meeting*, 693–704. Washington, DC, October 31 – November 4, 1999.
24. Yates, J., and W. Orlikowski. 1992. Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review* 17: 299–326.