INTERNET GENRES

Kevin Crowston

Syracuse University School of Information Studies

348 Hinds Hall

Syracuse, NY 13244

crowston@syr.edu

Tel: +1 (315) 443–1676

Fax: +1 (866) 265–7407

This research was partially supported by NSF IIS Grant 04-14482.

INTERNET GENRES

ABSTRACT

Rhetoricians since Aristotle have attempted to classify communications or documents into categories or "genres" with similar form, topic or purpose. This article surveys research on genre as it relates to Internet documents. The article briefly presents the concept of genre in general, and then reviews the evolution and emergence of genres on the Internet. It concludes with an examination of the possible use of genre for improving information access on the Internet, with specific discussion of the issues in developing taxonomies of genre and automatically recognizing document genre.

Keywords: Genre, World-wide web, Internet, Digital documents, Information access

Word count: 6908 words plus abstract, tables and references

INTERNET GENRES

INTRODUCTION

Rhetoricians since Aristotle have attempted to classify documents into categories or "genres" with similar form, topic or purpose. (In this article, we adopt a broad definition of document as "signifying objects"(*1*), meaning something that serves as evidence, regardless of the particular medium or form.) Numerous definitions of genre have been debated in the applied linguistics community (e.g., *2, 3-7*), while other groups have struggled with similar notions, such as discourse or document types, e.g., in SGML (*8*). This article, for example, is an instance of the *encyclopedia article* genre, commonly used to communicate the state of knowledge in a field. Other common genres include *letters* and *memos*, *project team meetings* and *TV sitcoms*, all immediately recognizable by their typical purpose and characteristic form. In this article, we provide a brief review of the concept of genre in general, drawing primarily on the cultural perspective of the New Rhetoric school (*4, 9*), before addressing the theory and applications of genre for digital documents found on the Internet.

Genres define a class of similar documents. Such a categorization can be made on different bases and approaches to defining genre have focused on different aspects of documents (see *10* for a review of the development of the concept). Some genres are defined primarily in terms of the physical form, such as a *booklet* or *brochure*; still others in terms of the document form, such as *lists* or *directories*; while other genres are defined by purpose or function, such as a *proposal* or *inquiry*. However, most genres imply a combination of purpose and form (*2*), such as a *newsletter*, which communicates "the

news of the day", including multiple short articles and is distributed periodically to subscribers or members of an organization.

There is a close relation between the terms "genre" and "text type". Lee (*11*) distinguished between them by suggesting that the term "genre" is often used for definitions based on external criteria, such as purpose, while "text type" is used for definitions based on internal criteria, such as form, but goes on to conclude that genre can be used "to describe most of the corpus categories we have seen". (Moessner (*12*) makes the same distinction.) For the purpose of this article, we will use the term genre as including aspects of both form and purpose.

Campbell and Jamieson (*2*) suggested that genres arise as responses to recurrent communicative situations. Given a recognized need to communicate (i.e., a purpose, *13*), individuals will typically express similar social motives, themes and topics in a communication with similar physical and linguistic characteristics (i.e., form), that is, they will communicate in a recognized genre. Miller (*6*) argued that the situations that give rise to the use of genres are social situations and that the process of creating genre is one of typification, as similarities in situations are recognized. If the typification is useful, the use of the genre becomes routine. She further argued that genres must accomplish a recognized social function and criticized the labeling of *environmental impact statements* as a genre because they did not meet this criterion, but rather had multiple conflicting motives.

Swales (*7*) similarly viewed genres as "a system for accomplishing social purposes by verbal means". He suggested analyzing documents in terms of their exhibited characteristic moves, defined as "a functional unit, used for some identifiable

rhetorical purpose" (*14*). For example, Swales (*7*) analyzed the *introduction, methods, results, discussion* and *conclusion sections* of research articles as well as briefly touching on other research process genres, such as *abstracts*, *presentation*, *theses* and *grant proposals*. Following Swales's (*7*) approach, researchers have described other scientific genres such as *research paper abstracts* (*15*) and *discussion sections* (*16*), *grant proposals* (*14*), *posters* (*17*) and other *visual presentations at scientific meetings* (*18*) and *French thesis defences* (*19*). Researchers have examined as well as journalistic genres such as *news* (*20, 21*), *editorial letters* (*22*) and *magazine covers* (*23*) and genres from the commercial world, such as *business letters of negotiation* (*24*), *letters of application for jobs* (*25*), *Chinese sales letters* (*26*), *environmental impact statements* (*27*), *building reviews* (*28*) and *philanthropic direct mail* (*29*).

THEORY: DOCUMENTS AND GENRE

Viewed from the perspective of the reader of a document, identification of a document's genre makes the document more easily recognizable and understandable, thus reducing the cognitive load of processing it (*30*). Because we drew on the *encyclopedia article* genre in writing this article, for example, a reader should be able to more quickly determine the purpose and content of our communication and begin to evaluate its contribution. The form of the article gives hints to its meaning and appropriate uses. Similarly, following the form of the genre of a *research article*, *letter of application*, *parking ticket* or *bank statement* makes the corresponding documents easier to recognize, to assess and to use. This effect has been demonstrated experimentally: Vaughan and Dillon (*31*) found that readers of a Web newspaper that followed genre conventions reported better comprehension, usability and ease of navigation than

readers of a purported newspaper that did not. Knowledge of genres can similarly help creators of documents by providing a known form for achieving a communicative purpose. Rather than having to innovate in all aspects of a document, a writer can reuse the form of a familiar genre to achieve their purpose. Teaching of genres is thus seen as an important aspect of teaching language and communication, e.g., in elementary (*32*), second language (*33*) and English for Special Purposes (*34*) education.

In addition to helping senders and recipients singly, knowledge of genres can also facilitate interaction, since multiple communications may be performed in a recognizable pattern, what Bazerman (*35*) called a "genre system". In other words, recognizing that a communication is of a particular genre may suggest the form expected for the reply. Swales (*36*) similarly defined a "genre chain" as a sequence of documents of different genres on a given topic. Examples include the sequence of *examination* and *cross-examination* in a trial, or the cycle of *article* submission to a journal or conference, reviews, *final acceptance* or *rejection letters* and publication. In each case, knowledge of the genre system provides information about how to proceed with the interaction. For example, Tardy (*37*) examined the genre system of research funding, identifying an interlocking set of genres such as a *research agency's mission statement* and *program information*, informal *meetings with program officers*, *applications*, *panel reviews* and *decisions*. Antunes, Costas and Dias (*38*) analyzed genres related to electronic meetings. Features of a genre may enable their use in a genre system: for example, page numbers in a *technical paper* make it possible to more precisely cite concepts or quotations from the paper, thus binding the paper into the literature.

For genres to be of aid in communication though, they must be shared across the members of particular discourse communities (*7*). Thus, genres are socially situated.

Teachers and rhetoricians have acknowledged that learning about genres, and through genres, is learning how to participate in a community (e.g., *32*). On the other hand, a genre may be unfamiliar or hard to understand for someone outside of the community. In fact, recognition of a particular genre is one sign of membership in a particular community (one needs only to witness the uncritical citing habits of college students who are not able to distinguish a popular Web genre from a serious scholarly one). Indeed, Freedman and Medway (*39*) suggest that incomprehensible genres may even be used deliberately to defend positions of privilege. Thus genres are not only a type of discourse but are themselves embodiments of social or communicative events (*40*).

Leveraging the situated character of genres, Yates and Orlikowski (*41, 42*) proposed using them as a lens into organizational practices. Studying organizational discourse as examples of genres anchors and situates the investigation in a way that other approaches could not because each genre has within it the information for how it should be interpreted within a particular discourse community. For example, Yates, Orlikowski and Okamura (*43*) contrasted the explicit and implicit structuring of genres in the introduction of a new communications system, and Schryer and Spoel (*44*) examined genres as a way to understand professional identity formation of medical students and of midwives.

To capture these social elements, Orlikowski and Yates (*42*) introduced the notion of a "genre repertoire", that is, the set of genres in use within a community. They noted that different communities use different genres in their communication, and use common genres with different frequencies. (Hengst and Miller (*45*) refer to this situation as the "pervasive heterogeneity" of genres.) These differences provide one source of insight into the communicative (and other) practices of the community. For

example, a community of social scientists and computer scientists can be distinguished by the frequency of use of different paper genres, as well as the paucity of computer programs and program documentation created in the former, reflecting different modes of research.

Of course, the reality of documents and genre is not as clear-cut as the theory above suggests. First, not all communications or documents are necessarily generic. For example, Swales (*7*) ruled out conversation as a genre, arguing that the concept only applies to completed texts. Pieces of text or unfinished texts may not be generic. As well, if genres are formed in response to recurrent communicative needs, then communications in novel (or even relatively) novel situations may also not be generic. This circumstance may describe large portions of the communication taking place via the Internet (and so of the documents found there). Even in typified circumstances, an author may deliberately or inadvertently fail to reproduce the necessary form of a genre, again producing a non-generic document. Finally, the relation amongst genres in a genre system is rarely linear; in any given situation, there may be a set of appropriate responses (*34*).

Furthermore, the boundaries between genres are often fuzzy and documents may show considerable variation in form even within a genre (*12, 46*). As a result, it may be helpful to think of genres defined by exemplars and documents as being more or less good examples of a genre rather than attempting to draw firm boundaries. As well, as Orlikowski and Yates (*42*) point out, some communications use multiple genres simultaneously, such as a *proposal* embedded in a *memo* or a *book chapter* in a *book*, along with *table of contents* or *index*. Other documents might have some mixed attributes, e.g., *bibliography* that is also an *index to a document collection*. As a result, the mapping of

documents to genre may be one to many. This mixing of genres is likely to be particularly problematic for digital documents, which can have multiple forms of presentation.

Further complicating the study of genre is the fact that knowledge of genres tends to be quite tacit and situated in use. Being able to recognize or use a genre is typically knowing "how to go on" (*47*) in a particular situation rather than formal articulable knowledge. The genre analyses cited above are research contributions precisely because they required research to uncover the hidden regularities in the documents. In practice, an individual may recognize and be able to use a document, while being unable to articulate a name for its particular genre or to say how they know what it is or which particular features are important for defining the form. When tested out of context, users may find genre hard to apply. For example, Santini (*48*) had 135 users classify the genre of 25 Web pages and found high levels of agreement on only a handful of documents, while others attracted as many as four equally popular labels. Similarly, in a study with 102 webpages, Rosso (*49*) found a wide variety of genre terms (49) when three users were asked to pick their own labels. On the other hand, in a study with 257 users and 55 pages, he obtained a reasonable level of agreement (70%) when the choice was restricted to one of 18 terms, again suggesting that genre can be recognized more easily than produced.

INTERNET GENRES

We now turn our focus to genres of digital documents found on the Internet. The Internet and in particular the World-Wide Web provides a particularly interesting setting in which to study the use and development of genres for several reasons. First,

the capabilities of the new medium have led to the development of many new genres of communication. Space limitations preclude a comprehensive review of the development of the Internet from the ARPAnet's four computers in 1969 to the immense and nearly ubiquitous network of today, but we note that the increased functionality of the Internet has been paralleled by an explosion in Internet genres. In particular, the technology of the Web extends the notion of a document—and thus the notion of genre—because Web pages can provide functionality in addition to information. Indeed, some Web pages are more comparable to computer interfaces than to conventional paper documents. As a result, functionality may be important in understanding genres on the Internet. Furthermore, the rapid development of this medium suggests a high level of experimentation with potential genres. Bearman (*50*), for example, notes the rapid evolution in what he refers to as "forms of material" in electronic media in general.

Second, since many Web sites are open to the public, many examples of Web communication are easily available for study. Furthermore, because there is no central management of the Internet or the Web, there is no explicit management or enforcement of genres of communication, as might happen in the introduction of a communication system in a corporate environment (*51*). Instead, individual Web developers individually choose how to present their information, drawing on their understanding as members of a community, what Orlikowski et al. (*51*) called implicit structuring (in this case, from the point of view of the Web page developer rather than the recipient of the communication). Yates and Sumner refer to the "democratization of genre production" as "communities evolve increasingly well-defined genres to better support their particular communicative needs and work practices" (*52*). However, even in this

free-for-all, mutual acceptance of genre is important to enable communications. Yates and Sumner (*52*) argue that on the Web, genres help in both the production and consumption of documents because genre adds "fixity" in a medium that does not otherwise distinguish very well between text types (say, a *book* and a *post-it*).

Finally, there are many communities meeting on the Web, bringing experiences with different genres and using the Web for many different purposes. The Web is sometimes used for direct communication where someone with a Web server "delivers" a document to members of a known community by giving them a URL. For example, some academics use the Web to communicate with colleagues by publishing their own papers, and with students by publishing syllabi and assignments. Another example of communication within a predictable community is computer companies announcing new products, publishing catalogs or providing troubleshooting tips on-line for their customers. However, in many other cases the audience is unpredictable. Unlike the Usenet or electronic mail groups, there is no clear separation of communities into different channels of communication (as is the case for journals or talks given at conferences, for which the audience is likely to have shared interests).

Indeed, it seems a stretch to say that there is a single Web community at all. Instead, the genre repertoire reflected in a collection of Web pages will be the result of interactions within and among multiple communities. In some cases, a genre may act as a type of boundary object (*53*), providing a common point of contact between different groups (*39*). In others, this mixing may lead to genre confusion, meaning that there is a practical need to understand the way genres enable communication. For example, organizations have used the Web to publish information such as *product brochures*, *annual reports*, country, state, and city *home pages*, *government agency press releases*, etc.

These organizations tend to use familiar genres when putting information on the Web. However, a person happening to reach a document on one of their Web sites has a good chance of being outside the community in which that genre evolved. As a result the document may be confusing and the communicative purpose lost.

Evolution of Genres

To understand the evolution of genres on the Web, we draw on studies of how pre-digital genres have evolved over time. Drawing on Giddens' (*47*) structuration theory, Orlikowski and Yates (*42*) argued that, "People produce, reproduce and change genres through a process of structuring". As members of the community draw on their knowledge of a genre repertoire to communicate, they reinforce the use of these genres, making them more appropriate or legitimate for use in the given situation. For example, by creating an order entry Web page that draws on the genre of a paper order form, a designer reinforces the appropriateness of the order form genre for this type of communication, making its use in future situations more likely. In other words, the set of genres in use (i.e., the genre repertoire) is both a product of and a shaper of the communicative practices of a community.

Orlikowski and Yates (*42*) suggested that in a new situation (such as in the introduction of a new medium such as the Internet) individuals will typically draw on their existing genre repertoires, reproducing genres they have experienced as members of other communities. For example, Görlach (*46*) notes the emergence of new genres, such as the *cooking recipe*, that draw on early genres, in this case the *medical recipe* (though he uses the term text type rather than genre). However, people are also free to modify a genre and communicate in a way that invokes only some of the expected

aspects of a form. If these changes become repeatedly used (i.e., typified), they too may become accepted and used together with or instead of existing genres, thus extending or altering the genre repertoire. Because the definition of genre relies on social acceptance, it is impossible to define the exact point at which a new genre emerges from the old one. Acceptance may take many years. However, after some period of coexistence, the new combination of form and purpose may become generally recognized and even named as a separate genre. As well, genres may be accepted in different communities at different rates. The emergence of distinctive new genres would be one sign of the formation of a new community with new communicative practices.

Take for example the *academic journal article*, a distinctive genre with the communicative goal of reporting research results and establishing a researcher's credentials and reputation. Journal articles have moved nearly intact to the Web and can be found in many online databases and on publishers' Web sites, as well as on the Web sites of individual authors. These documents are often identical to the paper versions, an example of a reproduced genre (literally reproduced, as in many cases, the Web versions are simple scans of the paper form). However, in a few cases, the form of the journal article has begun to change to take advantage of the possibilities of linking or embedding information. For example, citations in papers may be hyperlinks to the referenced articles. In part to enable easier searching of results, some journals now require *structured abstracts*, a particular genre of *abstract* with a distinctive form. Some publisher sites allow users to comment on papers, enabling more interactive follow-up discussion. Such documents are examples of adapted genres. Even more interesting is the emergence of forms other than journal articles for reporting research results, such as

datasets, *software* or other products of research, all examples of emergent genres. For example, researchers in genetics may publish *gene sequences* in a variety of specialized databases that can substitute in some case for a journal publication (*54*). Similar efforts are underway in other disciplines, meaning that the eventual form used for reporting research may bear only passing resemblance to the self-contained 20–25 page articles of today.

Related changes are already visible at the level of *journals*. For example, while paper journals are often reproduced on a Web site, there are now journals that publish only on-line editions, which need not conform to typical page limits or even have volumes at all (an adapted genre). There are suggestions that the increased use of online databases for finding articles (either by topical search or citation indexing), as opposed to subscribing to a handful of specific journals, is blurring previously distinct publications into "the literature" (*55*). As a result, the need for individual journals with distinct missions and readerships may be reduced, as is the utility of the genre itself. Similarly, the widespread use of *working paper* archives and of general search engines that cover such "grey literature" (*working papers*, *conference papers*, etc.) may blur the boundary between these genres and journal articles and thus lead to hybrid or novel genres. (Indeed, such a shift may have already happened in disciplines that value conference papers on par with journal articles, reducing the distinction between the purposes of these genres.)

These emergent genres may be immediately accepted or, more likely, there may be a transition period during which the limits of the genre are renegotiated. For example, the *electronically distributed journal article* is still in transition (*56, 57*). It is being used, but this adapted genre is not yet completely accepted or considered legitimate for

all purposes (e.g., as evidence for a tenure case) by the academic community as a whole. As well, modifications of genres that are parts of genre systems may require corresponding changes to the rest of the system. For example, changes in citation habits will be necessary before page numbers can be dropped from the *technical paper* genre. Such interdependencies between genres will tend to slow the adoption of a new genre.

Internet Genres

While many genres have been reproduced more-or-less faithfully, as in the example of *journal articles*, the Internet has seen the rise of a few novel genres. In a study of 1000 Web documents, Crowston and Williams (*58*) were able to identify documents of many familiar genres and of a few genres that seemed to be new to the Web, such as the *hotlist*. A specific example of a novel genre is the *home page*. As early as 1996, Furuta and Marshall (*59*) noted the emergence of this genre as a result of the specific affordances of the Internet. While Bates and Lu in 1997 suggested that the *home page* genre was still inchoate (*60*), soon after Dillon and Gushrowski (*61*) found that the features of *personal home pages* seemed to have stabilized rather quickly. The *FAQ* (*Frequently Asked Questions*) document emerged as a distinct genre on the Usenet and was then translated to the Web. An AltaVista search done by Crowston and Williams (*58*) indicated approximately 170,000 Web pages with FAQ or "Frequently asked questions" in their title (a search in 2007 with Google finds more than 16 million such pages).

Other authors have identified a variety of Internet specific genres. Some have to do with particular communications media used on the Internet, such as *email message* or *Weblogs*. Gains (*62*) analyzed small collections of *business* and *academic email messages*. He

suggested that email as a whole has too many uses to be considered one genre and that the business emails analyzed "appear to follow the normal conventions for standard written business English", suggesting that they did not include new genres. On the other hand, he noted that *academic email messages* had some conversational features and suggested that the collection might include new genres of communications. Gruber (*63*) analyzed contributions to two *academic mailing lists* as forming a single genre with features of *academic letters* and *scholarly publication* as well as of oral communication. (It should be noted that these analyses were carried out when the medium was still novel to many users.) More recently, Barron (*64*) analyzed the properties of *unsolicited commercial email* (UCE or "spam"). She noted similarities between *spam* and other promotional genres, with moves such as capturing attention, establishing credentials, introducing the offer and multiple moves soliciting a response. She noted differences though, such as the need to capture attention to get the message opened and the inclusion of offers to unsubscribe the recipient.

A more recent innovation is the *Weblog* or *blog*. Based on an analysis of a random sample of 203 blogs, Herring et al. (*65*) characterized *blogs* as a "hybrid genre that draws from multiple sources" both offline and Internet. They distinguished several types of *blogs* with different forms and purposes as well as distinct origins. Most of their sample they characterized as *journal blogs*, which they suggested draw on the form and purpose of *diaries*, since they report the writer's feelings rather than other content. A few were *filter blogs*, providing pointers to other content of interest and perhaps deriving from the earlier Web genre of *hotlists* or, they suggest, from the offline antecedent of *letters to the editor*. Other types of blogs identified include *community*, *travel*, *memory* and *communications blogs*, though these were rare in their sample.

Other authors have analyzed particular genres of documents commonly appearing on the Internet, such as the *personal home pages* mentioned above (*59-61, 66*), *Web resumes* (*67*), *Internet advertising* ("netvertising") (*68*) and *online encyclopedias* (*69*). Howard (*70*) went as far as to claim the existence of a "Web vernacular", a particular form of Web page that was expected and so copied by commercial developers.

Still others have examined non-textual documents, such as *audio loops* (short sound sample) (*71*), *multimedia* (*72*) and *databases* (*73*), as well as mixed media such as *PowerPoint presentations* (*74*). Exploration of genres of digital documents therefore blends at some point into the study of non-textual genres. The discussion above can be extended to such documents. For example, we can see the recreation of genres of music (e.g., *pop song* or *concerto*), even as the medium shifts from LP to CD to MP3, as well as adaptation of the genres of *televisions shows* to take advantage of the increased functionality (and decreased screen size) as they move from broadcast TV to Web distribution.

Genre classification

While we do not have a list of all of the genres on the Internet (nor is it clear that such a thing is even possible), there is a substantial body of work categorizing genre in printed documents and some work studying them on the Internet. Many attempts to develop a categorization of genres have been top-down, that is, they analyzed a set of documents based on theoretical principles or according to *a priori* classifications. A key difference in these efforts is the number of genre categories distinguished. Many studies of Web pages have used fewer broader categories: for example, zu Eissen and Stein (*75*) used only 8 genres (*help; article; discussion; shop; portrayal, non-private; portrayal, private;*

link collection; and *download*). At the other extreme, Görlach (*46*) offered a catalog of some 2000 genre (or text type) terms, which is intended to be an exhaustive list of the terms used in English. Somewhere in between, Lee (*11*) categorized documents in the British National Corpus (BNC) into 70 genres or subgenres (with some document assigned more than one genre). However, he notes that the genre terms used were "meant to provide starting points, not a definitive taxonomy", for example grouping *textbooks* and *journal articles* as *academic texts* that can be further distinguished by medium.

If the classification includes more than a handful of terms, it is useful to group together similar terms and necessary to deal with terms of different levels of generality (*11*). For example, *social science papers* might be grouped with *computer science implementation papers*, *biology research papers* and so on as examples of *academic papers*. These genres share some similarities, such as a title, abstract and bibliography, but differ in other particulars, such as the expected section headings, types of arguments or evidence. Many organized lists of genres are structured as single hierarchies. Figure 1 shows a small section of the hierarchy of genres of Web documents identified by Crowston and Williams (*58*). *Advertisements* and *announcements* are both examples of *declaratory document* genres; *classified advertisements* are a special kind of *advertisement*, and so on. Similarly, *social science papers*, *computer science papers* and *biology papers* might be seen as examples of a more general genre of *research papers*, which are in turn examples of *papers* or *articles*.

An advantage of a hierarchy is that it avoids the need to predetermine the level of detail needed in the classification. Depending on the circumstances, we can consider genres at any of these different levels and different levels might be more or less useful

for different purposes. Of course, there is no guarantee that convenient and well-known terms will exist for all levels of the hierarchy. A second criticism of traditional hierarchies is that they rely on a single organizing principle, which may not be useful or appropriate for all cases. Harrell and Linkugel (*5*) note that there are multiple bases on which such a classification could be constructed. To overcome this problem, Kwaśnik and Crowston (*76*) suggested using the facetted-analysis approach, following the example of previous genre-identification studies such as Päivärinta (*77*), Tyrväinen and Päivärinta (*78*) and Karjalainen et al. (*79*) who looked at the management of enterprise documents, and Kessler, Nunberg and Schuetze (*80*) who sought to identify a limited set of facets for communicative purposes. Crowston and Williams (*58*) based their classification of genres on the *Art and Architecture Thesaurus* (*81*), which is also a facetted classification overall.

The previously discussed classification schemes can be described as top-down or *a priori*. Though this approach is quite common in conventional settings, given the communal nature of genres, a top-down approach to classification seems problematic. As genres are socially constructed, different social groups using documents with similar structural features may think about them and describe them differently (i.e., as different genres, though perhaps similar text types *11*). While many genres may be widely shared, even more will be local to particular groups. Documents may be of genres that are not necessarily vetted by traditional schemes, particularly documents that come out of domain-specific work. Furthermore, while genres have always been conceptualized as dynamic, Dillon and Gushrowski (*61*) point out that genres are no longer necessarily "slow-forming, often emerging only over generations of production and consumption…". Thus, a static typology of genre or document forms may not be

sufficient to describe the emerging and dynamic genres in use on the Internet. It seems important instead to capture users' own language and understanding of genres.

Some researchers have attempted to identify genres bottom-up through user studies. Dewe, Karlgren and Bretan (*82*) asked users to provide lists of genres found on the Web and received 67 responses. They noted though that users tended to conflate genre and topic. Nilan, Pomerantz and Paling (*83*) surveyed 242 Web users in person and via the Web about their purpose in searching the Web, the genre of document expected and the actual document found, and collected 1335 example pages; genres could be assigned for 1076, giving a total of 116 genres. They then grouped the genres given, first following the plan of the Art and Architecture Thesaurus and second based bottom-up on the user responses. The second classification was much broader and shallower than the first, and some genres were classified in multiple locations, suggesting that there may be difference between genres in use and formal structured knowledge. Freund, Clarke and Toms (*84*) developed a list of 16 genres based on interviews with users and analysis of document repositories in a particular organization. They found that the 16 genres covered about 75% of documents randomly sampled from the repositories.

APPLICATIONS OF GENRE IN INFORMATION SCIENCE

In this final section, we discuss how ideas about genre might be applied to system design. Researchers have addressed a variety of information systems design questions, such as the use of genre to design electronic meeting systems (*38*) or to guide assembly of output documents from a content management system, e.g., creating training documents by reusing content from operations manuals (*79, 85*). We will focus

on one particular question likely to be of interest and importance to information scientists, namely how genre metadata might be used to improve information access on the Internet.

A significant problem in information access is that topic alone is not enough to define an information problem. Different users may require different solutions for seemingly similar topics because the situation (or context) of the user determines not only what topics are requested and what strategies are invoked in searching and evaluating output, but also what types of resources are considered relevant and useful. For example, "methods for learning mathematics" (a topic) will be construed differently by a student, by a parent and by a classroom teacher because of their different information-use situations. Indeed, even the same user may require different information at different times.

Although we know that it is important to understand the situation of the user, the actual representation of situations and then their implementation in a system remains a difficult problem. Our efforts to create user profiles, universal situation grammars, and so on suffer from limitations of scope to specific domains and lack of extensibility and flexibility. However, in a study using TREC data, Karlgren (*86*) found significant differences in style between documents judged relevant and not relevant, suggesting that non-topical metadata can be helpful. In particular, inclusion of genre information as non-topical characteristics of the documents might be useful as a signal to their purpose and so fit the user information need. For example, a university professor looking for information about computer database systems for the class that she teaches would most likely be interested in documents of educational genres (e.g., *syllabi*, *assignments*, *class notes*). On the other hand, when working on a research paper in

the database area, the same professor would more likely appreciate scholarly work (e.g., *research papers*, *annotated biographies*, *calls for papers*). The relevant documents for these two searches would be quite different, even though the topic and query keywords might be nearly the same.

Explicit identification of genre seems likely to be particularly important for large digital collections (the Web being the largest) because—unlike earlier collections of documents comprising a limited set of genres (e.g., a document database containing primarily *journal articles*)—these collections contain documents addressing a diversity of discourse communities with a diversity of genres (e.g., *journal articles* but also *magazine articles*, *hot lists*, *memos*, *home pages*, *class syllabi*, etc.). A user searching such a diverse document collection by topic will likely receive some documents of relevant genres along with many documents of irrelevant genres—a low precision result—even if all retrieved documents conform to search specifications regarding the topical content of the document. This analysis suggests that one way to exploit genre information is to create specialized search engines that retrieve only documents of a particular genre. This approach has already been followed by several systems, such as Indeed for *job listings*, CiteSeer and Google Scholar for *academic articles* and various Google specialized searches for *blogs*, *books*, *business addresses*, *news articles*, *patents*, *product sales pages* and *source code* (but excluding those specialized by media, e.g., searches for images or video).

A second way to use genre information is to enrich queries with information about expected genres of the results, e.g., as a form of relevance feedback (*86, 87*). Because most genres are characterized by both form and purpose, identifying the genre of a document provides information as to the document's purpose and its fit to the

user's situation, which can be difficult to assess otherwise. When medical information is sought, identical keywords might retrieve a *newsletter*, a *personal home page*, a *journal article* or a hospital's *patient-information site*. A person searching for one of these genres is unlikely to be satisfied by the others.

Knowledge of the form of genres can also help in the process of matching documents to queries. For example, an *FAQ* document is divided into question and answer pairs (indeed, documents of many genres have a form with repeated subpieces). Requiring search terms to be found in the same question-answer pair (or subpiece) may reduce spurious matches or false drops.

Once a search is completed, document genre may be useful to improve the accuracy of relevance judgments made to rank order search results. It has been noted that some genres are less likely to be relevant for the majority of search tasks. This implies that certain Web pages could be promoted or demoted in the ranked results if their genre were known. For example, it has been claimed that most searchers are not interested in retrieving *personal home pages* (*88*), so the latter could be moved down the results list by request. Bretan (*89*) suggested using genre to group search results while Freund, Clarke and Toms (*84*) suggest filtering results.

Finally, a one-size-fits-all approach to summarizing or evaluating Web documents is likely to misrepresent many documents when confronted with diverse genres. For example, a *newspaper article* can often be summarized by the first few paragraphs of the document, but such an approach will not work for a *home page* or *FAQ* (*90*). Rehm (*91*) analyzed the necessary components for a document to be an *academic home page* as a prelude to extracting information for further processing.

Automatic Identification of Document Genre

For genre to be useful in large-scale systems, techniques will be needed to automatically assign genre to large collections of documents. Researchers have tried numerous approaches to automatic genre identification; a sampling of studies is shown in Table 1. This work has used statistical (e.g., regression or discriminant analysis) or machine learning techniques (e.g., decision trees or support vector machines) to classify documents into genres based on features in the documents (Sebastiani (*92*) has reviewed these techniques). Comparing the algorithm's assignment to the known genre measures the success of the classification. Table 1 presents accuracy, though the some authors compute both precision and recall. For example, Freund, Clarke and Toms (*84*) suggest that for filtering, higher recall might be preferred even at the cost of lower precision, to ensure that all documents of a given genre are presented to the user.

Document features that have been suggested for use in genre recognition include counts of specific words or of closed-class words (e.g., days of the week), counts and ratios of parts of speech, word and sentence length, layout features (*93*), punctuation (*94*), URLs and HTML tags (*95*) and even the level of spelling and typing errors in the document (*96*). There has been a trend towards the use of larger and more comprehensive feature sets, from which the machine learning techniques can pick useful subsets. Which features are useful and necessary is still a topic for research. For example, Dewdney et al. (*93*) suggested that presentation features alone were sufficient to recognize genre, while Ferizis and Bailey (*97*) found that POS tagging was not needed for genre recognition; since it is computationally expensive, avoiding these features may be desirable. Dong et al. (*98*) found that more features increased recall but decreased precision, and that including more types of features improved both.

For statistical analysis or machine learning to be successful requires a large set of categorized documents to use as a training set. Creating such corpuses is time-consuming, so early studies used preexisting ones, such as the Brown corpus, which includes 500 samples (802 documents) in 15 text categories (*press reportage; press editorial; press reviews; religion; skill and hobbies; belles letters; miscellaneous US government and house organs; learned; fiction general; mystery and detective fiction; science fiction; romance and love story;* and *humour*). However, this corpus is not very suitable for the purpose as the average of 30 samples per genre is small for training and the categories mix topics and genres. More recent studies have created corpuses that are more focused and often larger (as large as 5–10 thousand pages with hundreds of examples of each genre). A concern here is the diversity of the documents included. Techniques applied to a corpus that includes only particular genres will yield more precise results than would be obtained with a more diverse sample.

Another key issue is number of genres to be recognized. Most studies have used small number of categories (from 4 to a maximum of 32, with a median of 10 in the studies in Table 1), which is small compared to the hundreds found in the user studies and general Web surveys reviewed above. Experimenters have examined fewer genres because increasing the number to be detected reduces precision and demands a larger corpus for training. But as Boese (*99*) noted, definitions of broader categories have be "softer" to include more documents and so very broad categories may be less useful to users or for the information access systems discussed above. Another concern is that the design of classification interacts with the features chosen. Grouping together documents with similar purposes but dissimilar structural features (i.e., genres as opposed to text types) may be useful for users, but create problems for automatic classification. With a

facetted classification, different techniques could be used recognize individual facets (*100*). Researchers have experimented with recognizing speech acts in email (*101*), a document's degree of expertise, detail and subjectivity (*102*) or positive and negative tone in reviews (*103*).

CONCLUSIONS AND OUTLOOK

The research reviewed in this article suggests that genre provides a useful lens for examining Internet documents. First, knowing the set of genres in use by a target audience can help ensure that information presented is easily understood and used. As well, there seems to be good potential to incorporate genre meta-data in information access systems. However, to realize these benefits, more research is needed. First, we do not as yet have a fully articulated set of data that reveals what genres various target groups recognize nor for what tasks they find documents of specific genres useful. Second, while genre recognition has improved, it is still limited in the number of genres that can be detected reliably. These two research agenda interact, as better knowledge of the genres in use will inform our attempts to recognize and use them in future Internet systems.

REFERENCES

1: Buckland, M. K. What is a "document"? *Journal of the American Society for Information Science* **1997**, *48* (9), 804–809.

2: Campbell, K. K.; Jamieson, K. H. Eds., *Form and genre: Shaping rhetorical action* (Speech Communication Association, Fall Church, VA, 1978)

3: Johns, A. M.; Bawarshi, A.; Coe, R. M.; Hyland, K.; Paltridge, B.; Reiff, M. J.; Tardy, C. Crossing the boundaries of genre studies: Commentaries by experts. *Journal of Second Language Writing* **2006**, *15* (3), 234-249.

4: Hyon, S. Genre in three traditions: Implications for ESL. *Tesol Quarterly* **1996**, *30* (4), 693-722.

5: Harrell, J.; Linkugel, W. A. On rhetorical genre: An organizing perspective. *Philosophy and Rhetoric* **1978**, *11*, 262–281.

6: Miller, C. R. Genre as social action. *Quarterly Journal of Speech* **1984**, *70*, 151–167.

7: Swales, J. M. *Genre Analysis: English in Academic and Research Settings*; Cambridge University Press: New York, 1990.

8: Dollar, C., in *Conference on Electronic Records in the New Millennium*. (Vancouver, BC, 1994) pp. 25–38

9: Muntigl, P.; Gruber, H. Introduction: Approaches to Genre *Folia Linguistica* **2005**, *39* (1–2), 1–18.

10: Breure, L. (Information and Computing Sciences, University of Utrecht, Utrecht, The Netherlands, 2001), vol. 2007,

11: Lee, D. Y. W. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* **2001**, *5* (3), 37–72.

12: Moessner, L. Genre, text type, style, register: A terminological maze? *European Journal of English Studies* **2001**, *5* (2), 131–138.

13: Bitzer, L. F. The rhetorical situation. *Philosophy and Rhetoric* **1968**, *1*, 1–14.

14: Connor, U.; Mauranen, A. Linguistic Analysis of Grant Proposals: European Union Research Grants. *English for Specific Purposes* **1999**, *18* (1), 47-62.

15: Martín, P. M. A genre analysis of English and Spanish research paper abstracts in experimental social sciences. *English for Specific Purposes* **2003**, *22* (1), 25-43.

16: Peacock, M. Communicative moves in the discussion section of research articles. *System* **2002**, *30* (4), 479-497.

17: MacIntosh-Murray, A. Poster presentations as a genre in knowledge communication: A case study of forms, norms and values. *Science Communication* **2007**, *28* (3), 347–376.

18: Rowley-Jolivet, E. Visual Discourse in Scientific Conference Papers: A Genre-Based Study. *English for Specific Purposes* **2002**, *21* (1), 19-40.

19: Maingueneau, D. Analysis of an academic genre. *Discourse Studies* **2002**, *4* (3), 319-342.

20: Vestergaard, T. That's not news: Persuasive and expository genres in the press. In *Analysing Professional Genres*; A. Trosborg, Ed.; John Benjamins: Amsterdam, 2000, 97–119

21: Ljung, M. Newspaper genres and newspaper english. In *English Media Texts-Past And Present. Language And Textual Structure*; F. Ungerer, Ed.; John Benjamins: Amsterdam, 2000, 131-149

22: Flowerdew, J.; Dudley-Evans, T. Genre analysis of editorial letters to international journal contributors. *Applied Linguistics* **2002**, *23* (4), 463-489.

23: Held, G. Magazine covers: A multimodal pretext-genre. *Folia Linguistica* **2005**, *39* (1–2), 173–196.

24: Pinto dos Santos, V. B. M. Genre analysis of business letters of negotiation. *English for Specific Purposes* **2002**, *21* (2), 167-199.

25: Henry, A.; Roseberry, R. L. A narrow-angled corpus analysis of moves and strategies of the genre: "Letter of application". *English for Specific Purposes* **2001**, *20* (2), 153-167.

26: Zhu, Y. Rhetorical moves in Chinese sales genres, 1949 to the present. *Journal of Business Communication* **2000**, *37* (2), 156–172.

27: Bazerman, C.; Little, J.; Chavkin, T. The production of information for genred activity spaces: Informational motives and consequences of the environmental impact statement *Written Communication* **2003**, *20* (4), 455–477.

28: Caballero Rodriguez, R. Metaphor and Genre: The Presence and Role of Metaphor in the Building Review. *Applied Linguistics* **2003**, *24* (2), 145-167.

29: Upton, T. A. Understanding Direct Mail Letters as a Genre. *International Journal of Corpus Linguistics* **2002**, *7* (1), 65-85.

30: Bartlett, F. *Remembering: A Study in Experimental and Social Psychology*; University Press: Cambridge, England, 1932/1967.

31: Vaughan, M. W.; Dillon, A. Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies* **2006**, *64*, 502–526.

32: Chapman, M. L. Situated, social, active: Rewriting genre in the elementary classroom. *Written Communication* **1999**, *16* (4), 469-490.

33: Hyland, K. Genre-based pedagogies: A social response to process. *Journal of Second Language Writing* **2003**, *12* (1), 17-29.

34: Swales, J. M. Languages for specific purposes. *Annual Review of Applied Linguistics* **2000**, *20*, 59–76.

35: Bazerman, C. Systems of genres and the enactment of social intentions. In *Genre and the New Rhetoric*; A. Freedman, P. Medway, Eds.; Taylor and Francis: London, 1995, 79–101

36: Swales, J. M. *Research Genres: Exploration and Applications*; Cambridge University Press: Cambridge, 2004.

37: Tardy, C. M. A genre system view of the funding of academic research. *Written Communication* **2003**, *20* (1), 7–36.

38: Antunes, P.; Costa, C. J.; Pino, J. A. The use of genre analysis in the design of electronic meeting systems. *Information Research-an International Electronic Journal* **2006**, *11* (3).

39: Freedman, A.; Medway, P. Locating genre studies: Antecedents and prospects. In *Genre and the New Rhetoric*; A. Freedman, P. Medway, Eds.; Taylor and Francis: London, 1994, 1–22

40: Askehave, I.; Swales, J. M. Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics* **2001**, *22* (2), 195-212.

41: Yates, J.; Orlikowski, W. J. Genres of organizational communication: A structurational approach to studying communications and media. *Academy of Management Review* **1992**, *17* (2), 299–326.

42: Orlikowski, W. J.; Yates, J. Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly* **1994**, *33*, 541–574.

43: Yates, J.; Orlikowski, W. J.; Okamura, K. Explicit and implicit structuring of genres in electronic communication: Reinforcement and change of social interaction. *Organization Science* **1999**, *10* (1), 83–103.

44: Schryer, C. F.; Spoel, P. Genre theory, health-care discourse, and professional identity formation *Journal of Business and Technical Communication* **2005**, *19*, 249–279.

45: Hengst, J. A.; Miller, P. J. The heterogeneity of discourse genres: Implications for development. *World Englishes* **1999**, *18* (3), 325-341.

46: Görlach, M. *Text Types and the History of English*, Trends in Linguistics. Studies and Monographs 139; Mouton de Gruyter: New York, 2004.

47: Giddens, A. *The Constitution of Society: Outline of the Theory of Structuration*; University of California: Berkeley, 1984.

48: Santini, M. Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing and Management* **2007**.

49: Rosso, M. A., in *5th ACM/IEEE-CS Joint Conference on Digital libraries*. (Denver, CO, 2005)

50: Bearman, D. *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*; Archives and Museum Informatics: Pittsburgh, 1994.

51: Orlikowski, W. J.; Yates, J.; Okamua, K.; Fujimoto, M. Shaping electronic communication: The metastructuring of technology in the context of use. *Organization Science* **1995**, *6* (4), 423–444.

52: Yates, S. J.; Sumner, T. *Digital genres and the new burden of fixity* Paper presented at the Hawaiian International Conference on System Sciences (HICCS 30); Wailea, HA, 1997.

53: Star, S. L.; Griesemer, J. R. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. In *Social Studies of Science* Sage: Newbury Park, 1989; 19, 387–420

54: Krawczak, M.; Ball, E. V.; Fenton, I.; Stenson, P. D.; Abeysinghe, S.; Thomas, N.; Cooper, D. N. Human gene mutation database: A biomedical information and research resource. *Human Mutation* **2000**, *15* (1), 45–51.

55: Nunberg, G. The places of books in the age of electronic reproduction. *Representations* **1993**, *42* (Spring), 13–37.

56: Kling, R.; Covi, L. Electronic journals and legitimate media in the systems of scholarly communication. *The Information Society* **1995**, *11* (4), 261–271.

57: Harter, S. P. Scholarly communication and electronic journals: An impact study. *Journal of the American Society for Information Science* **1998**, *49* (6), 507–516.

58: Crowston, K.; Williams, M. Reproduced and emergent genres of communication on the World Wide Web. *Information Society* **2000**, *16* (3), 201-215.

59: Furuta, R.; Marshall, C. C. *Genre as Reflection of Technology in the World-Wide Web* (1996) (Hypermedia Research Lab, Texas A&M

60: Bates, M. J.; Lu, S. An exploratory profile of personal home pages: Content, design, metaphors. *Online & CDROM review* **1997**, *21* (6), 331–340.

61: Dillon, A.; Gushrowski, B. Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science* **2000**, *51* (2), 202–205.

62: Gains, J. Electronic mail–A new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes* **1999**, *18* (1), 81–101.

63: Gruber, H. E-mail discussion lists: A new genre of scholarly communication? *Wiener Linguistische Gazette* **1997**, *60/61*, 24–43.

64: Barron, A. Understanding spam: A macro-textual analysis. *Journal of Pragmatics* **2006**, *38* (6), 880-904.

65: Herring, S. C.; Scheidt, L. A.; Bonus, S.; Wright, E. *Bridging the gap: A genre analysis of weblogs* Paper presented at the the 37th Hawaii International Conference on System Sciences, 2004.

66: Killoran, J. B. The gnome in the front yard and other public figurations: Genres of self-presentation on personal Home Pages. *Biography-an Interdisciplinary Quarterly* **2003**, *26* (1), 66-83.

67: Killoran, J. B. Self-published Web resumes - Their purposes and their genre systems. *Journal of Business and Technical Communication* **2006**, *20* (4), 425-459.

68: Fortanet, I.; Palmer, J. C.; Posteguillo, S. The emergence of a new genre: Advertising on the internet (netvertising). *Hermes* **1999**, *23*, 93-113.

69: Emigh, W.; Herring, S. C. *Collaborative authoring on the web: A genre analysis of online encyclopedias* Paper presented at the the 38th Hawaii International Conference on System Sciences, 2005.

70: Howard, R. G. Toward a theory of the world wide web vernacular: The case for pet cloning. *Journal of Folklore Research* **2005**, *42* (3), 323–367.

71: Stillar, G. Loops as genre resources. *Folia Linguistica* **2005**, *39* (1–2), 197–212.

72: Lemke, J. L. Multimedia genres and traversals. *Folia Linguistica* **2005**, *39* (1–2), 45–56.

73: Manovich, L. Database as a genre of new media. *AI & Society* **2000**, *14*, 176–183.

74: Myers, G. Powerpoints: Technology, lectures, and changing genres. In *Analysing Professional Genres*; A. Trosborg, Ed.; John Benjamins: Amsterdam, 2000, 177-191

75: zu Eissen, S. M.; Stein, B. *Genre classification of web pages: User study and feasibility analysis* Paper presented at the the 27th Annual German Conference on Artificial Intelligence (KI 04); Ulm, Germany, 2004.

76: Kwasnik, B. H.; Crowston, K. *A framework for creating a facetted classification for genres: Addressing issues of multidimensionality* Paper presented at the the Hawai'i International Conference on System Science (HICSS), 5–9 January; Big Island, Hawai'i, 2004.

77: Päivärinta, T. *A genre approach to applying critical social theory to information systems development* Paper presented at the the 1st Critical Management Studies Conference, Information Technology and Critical Theory stream; Manchester, England, 1999.

78: Tyrväinen, P.; Päivärinta, T. *On rethinking organizational document genres for electronic document management* Paper presented at the the 32nd Annual Hawaii International Conference on System Sciences; Los Alamitos, CA, 1999.

79: Karjalainen, A.; Päivärinta, T.; Tyrväinen, P.; Rajala, J. *Genre-based metadata for enterprise document management* Paper presented at the the 33rd Annual Hawaii International Conference on System Sciences; Los Alamos, CA, 2000.

80: Kessler, B.; Nunberg, G.; Schuetze, H. *Automatic detection of text genre* Paper presented at the the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics; Madrid, 1997.

81: Petersen, T. *Art and Architecture Thesaurus*; Oxford: New York, 1994.

82: Dewe, J.; Karlgren, J.; Bretan, I., in *11th Nordic Conference of Computational Linguistics*. (Copenhagen, Denmark, 1998)

83: Nilan, M.; Pomerantz, J.; Paling, S. *Genres from the bottom up: What has the Web brought us?* Paper presented at the Asist 2001: the 64th Asist Annual Meeting, Vol 38, 2001, 2001.

84: Freund, L.; Clarke, C. L. A.; Toms, E. G. Towards genre classification for IR in the workplace In *ACM International Conference Proceeding*: Copenhagen, Denmark, 2006; 176, 30–36

85: Honkaranta, A., in *8th CAiSE/IFIP8.1 EMMSAD '03 workshop* (2003)

86: Karlgren, J. Stylistic experiments in information retrieval. In *Natural Language Information Retrieval*; T. Stralkowski, Ed.; Kluwer: Dordrecht, 1998

87: Roussinov, D.; Crowston, K.; Nilan, M.; Kwasnik, B. H.; Liu, X.; Cai, J., in *Thirty-Fourth Hawaii International Conference on Systems Science (HICSS–34)*. (IEEE, Maui, HI, 2001)

88: Chen, H.; Schuffels, C.; Orwig, R. Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation* **1996**, *7* (1), 88–102.

89: Bretan, I.; Dewe, J.; Hallberg, A.; Wolkert, N., in *WebNet '98*. (Orlando, 1998)

90: Marcu, D., in *14th National Conference on Artificial Intelligence (AAAI-97)*. (1997)

91: Rehm, G. *Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage* Paper presented at the the 35th Annual Hawaii International Conference on System Sciences, 2002.

92: Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys* **2002**, *34* (1), 1–47.

93: Dewdney, N.; VanEss-Dykema, C.; MacMillan, R. *The form is the substance: Classification of genres in text* Paper presented at the Workshop on Human Language Technology and Knowledge Management, ACL 2001 Conference, 6-7 July 2001; Toulouse, France, 2001.

94: Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Automatic text categorization in terms of genre and author. *Computational Linguistics* **2000**, *26* (4), 471–498.

95: Lim, C. S.; Lee, K. J.; Kim, G. C. Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management* **2005**, *41* (5), 1263-1276.

96: Stubbe, A.; Ringlstetter, C.; Schulz, K. U., in *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*. (Hyderabad, India, 2007) pp. 9–16

97: Ferizis, G.; Bailey, P., in *WWW 2006 Conference*. (Edinburgh, Scotland, 2006)

98: Dong, L.; Watters, C.; Duffy, J.; Shepherd, M., in *Hawai'i International Conference on System System (HICSS-41)*. (Kona, Hawai'i, 2008)

99: Boese, E. S. *Stereotyping the web: Genre classification of web documents* MS Thesis, Colorado State University (2005)

100: Kim, Y.; Ross, S., in *Hawai'i International Conference on System System (HICSS-41)*. (Kona, Hawai'i, 2008)

101: Carvalho, V. R.; Cohen, W. W., in *HLT-NAACL Workshop on Analyzing Conversations in Text and Speech (ACTS)*. (New York, NY, 2006) pp. 35–41

102: Dimitrova, M.; Finn, A.; Kushmerick, N.; Smyth, B. (2002)

103: Finn, A.; Kushmerick, N. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology* **2006**, *57* (11), 1506–1518.

104: Crowston, K.; Williams, M. Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society* **2000**, *16* (3), 201–216.

105: Karlgren, J.; Cutting, D. *Recognizing text genres with simple metrics using discriminant analysis* Paper presented at the the 15th International Conference on Computational Linguistics; Kyoto, Japan, 1994.

106: Wolters, M.; Kirsten, M. *Exploring the use of linguistic features in domain and genre classification* Paper presented at the the 9th conference of the European chapter of the Association for Computational Linguistics; Bergen, Norway, 1999.

107: Bisant, D. An application of neural networks to sequence analysis and genre identification. *International Journal of Pattern Recognition and Artificial Intelligence* **2005**, *19* (2), 199-215.

TABLES AND FIGURES

<declaratory document genres>	
advertisements	
classified advertisements	Short paid announcements appearing in a periodical sorted according to the good or service being offered or requested
announcements	Printed or published statements or notices that inform the reader of an event or other news
custom 404 page	A Web page announcing that the requested Web page could not be found on the server
news bulletins	
press releases	Official or authoritative statements giving information for publication in newspapers or periodicals

Figure 1. A section of a hierarchy of document genres (from *104*).

Table 1. Studies of automated classification of documents by genre.

Study	Techniques	Features	Corpus	Typology	Results
Karlgren & Cutting (*105*)	Discriminant analysis	20 linguistic features, such as adverb count, character count, first person pronoun count, type/token ratio	Brown corpus.	Brown corpus categories	52% accuracy overall; grouping all fiction together improved accuracy to 65%
Kessler, Nunberg & Schütze (*80*)	Logistic regression and neural nets	55 features including structural, lexical, character-level and derivative cues	Brown corpus.	Three part classification of 6 genres: reportage; editorial; scitech; legal; nonfiction; and fiction; plus "brow" (popular, middle, upper-middle and high) and narrative (yes or no)	"Good results" for reportage and fiction but not for other genres. Only small difference between using surface and structural cues
Dewe et al. (*82*)	Classification rules	40 features including lexical, textual and genre specific features	Own corpus of 1358 Web pages	11 genres: informal/private; public/commercial; searchable indices; journalistic material; reports; other running text; FAQs; link collections; asynchronous multi-party correspondence; and error messages	90% accuracy on first split; 66-75% on remaining decisions.

Study	Techniques	Features	Corpus	Typology	Results
Wolters & Kirsten (*106*)	k-nearest neighbour classification, RIBL; learning vector quantization; IBL	100, 500 or 1000 lemma features and 54 part of speech tags	LIMAS, a German corpus of 500 documents modelled on the Brown corpus	33 categories taken from the Deutsche Bibliographie (not all genres); but experiments run with fewer: academic texts from humanities and from science and technology; press texts; fiction; politics; law; and economy	75–100% precision in assignment, but each genre done separately noted that the corpus provided too little material for training
Stamatatos, Fakotakis & Kokkinakis (*94*)	Regression and discriminant analysis suggest that regression works better than other techniques with a small number of training instances	22 features: token-level measures (e.g., word counts) and outputs from NLP processing (e.g., number of noun phrases detected, average length of a noun phrase)	Own corpus of 250 modern Greek documents, 25 of each genre.	10 genres: press editorial; press reportage; official documents; literature; recipes; curricula vitae; interviews; planned speeches; and broadcast news, scripted.	82% accuracy overall. Press editorial and press reportage often confused.
Dewdney, VanEss-Dykema & MacMillan (*93*)	SVM, decision tree and naïve Bayes.	323 word features and 89 presentation features: closed-class words, parts of speech, word and sentence length, punctuation, layout features	CMU corpus of 9750 documents.	7 genres: advertisement; bulletin board; frequently asked questions; message board; radio news; Reuters newswire; television news	89% accuracy. Presentation features alone were sufficient to classify genre.

Study	Techniques	Features	Corpus	Typology	Results
Lim, Lee & Kim (*95*)	TiMBL	329 features in five sets: URL, HTML tags, token information, lexical information and structural information.	Own corpus of 1224 Web pages.	16 genres: Dewe et al.'s (*82*) taxonomy, plus product specifications and image collections, and splitting public and commercial homepages in two, and reports into research reports, official materials and informative materials	74% accuracy in assignment; improved slightly with optimal subset of features. Could not reliably determine genre for input pages and other.
Bisant (*107*)	Neural network with hidden layer, decision trees, SVM	89 features including part of speech and characters. Analysis shows 25 aren't useful	Own corpus of 5000 emails and webpages	10 genres: advertisement, business correspondence, data entry forms, e-mail administration, e-zine, friend correspondence, internet chat, news, notices and technical data	86% overall accuracy using neural nets with considerable variation. Decision trees 79% accurate. Notices confused with news.
Boese (*99*)	Bayes Net, decision trees, logit boost, bagging	Considered 1600 features: style, form, content. Narrowed to a set of 78.	Own corpus of 343 Web pages	Developed classification of 115 genres but used 10 for experiments: abstract; call for papers; FAQ; how-to; hub/sitemap; job description; resume/CV; statistics; syllabus; and technical paper.	91% accuracy using logit boost and 78 features.

Study	Techniques	Features	Corpus	Typology	Results
Freund, Clarke & Toms (*84*)	SVM light	"Bag of words", but no structural features.	Own corpus of 800 documents (about 50 per genre) drawn from repositories in a single organization.	16 genres (manuals; presentations; product documents; technotes, tips; tutorials and labs; white papers; best practices; design patterns; discussions/forums; cookbooks & guides; engagement summaries; problem reports; technical articles) developed within a specific organization.	81–97% recall, but lower precision.
Stubbe, Ringlstetter, & Schulz (*96*)	Decision trees	Hand selected sets of genre-specific features: form, vocabulary and parts of speech, complex patterns, level of typing errors.	Own corpus of 1280 Web pages, 40 of each genre.	32 genres in 8 broad classes: journalism, literature, information, documentation, directory, communication, nothing.	78% accuracy at first level; 72% at second level with considerable variation.
Dong et al. (*98*)	Naïve Bayesian	5, 20 or 100 features from form, content and functionality	Corpus of 1280 web pages; 170 of 4 genres (3 from (*48*)) plus 600 random pages	4 genres: FAQ; news; e-shopping; personal home pages	86-92% accuracy (precision). Precision improved by fewer features, but recall by more. More types of features improved both.