Genre Based Navigation on the Web

Dmitri Roussinov, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai and Xiaoyong Liu
Syracuse University, School of Information Studies
{droussin; mnilan; bkwasnik; jcai; xliu03}@.syr.edu; crowston@ist.syr.edu

Abstract

We report on our ongoing study of using the genre of Web pages to facilitate information exploration. By genre, we mean socially recognized regularities of form and purpose in documents (e.g., a letter, a memo, a research paper). Our study had three phases. First, through a user study, we identified genres which most/least frequently meet searchers' information needs. We found that certain genres are better suited for certain types of needs. We identified five (5) major groups of document genres that might be used in an interactive search tool that would allow genre-based navigation. We tried to balance the following dual objectives: 1) each group should be recognizable by a computer algorithm as easily as possible 2) each group has a better chance of satisfying particular types of information needs. Finally, we developed a novel user interface for a web searching that allows genre-based navigation through three major functionalities: 1) limiting search to specified genres 2) visualizing the hierarchy of genres discovered in the search results and 3) accepting user feedback on the relevancy of the specified genres.

1. Introduction

1.1. Grand Challenge of Information Access on the Web

Many researchers and practitioners consider Internet searching to be one of the most challenging areas for future research involving National Information Infrastructure (NII) application [1], [5]. Consumers use search engines to locate and buy goods or to make important decisions (such as choosing a vacation destination, medical treatment or election vote). Internet search technology may have economic, social, political, and scientific effects [10].

1.2. The Notion of Genre

The work described in this paper is based on the belief that recognition of the genre of a Web document can improve the quality of web searches. Rhetoricians since Aristotle have attempted to classify communications into categories or "genres" with similar form, topic or purpose. Numerous definitions of genre have been suggested in that community [2, 7, 13, 17]. In out study, we build on the definition of genre by Orlikowski and Yates [14] as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" (p. 543).

Some genres are defined primarily in terms of purpose or function, such as a proposal or inquiry; others 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. However, most genres imply a combination of purpose and form, 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. For example, this document is an example of a conference paper. It has a form familiar to many scientists: a title, authors and affiliations, an abstract, sections for background, work done, future plans, bibliography, and so on. More importantly, because genre implies both form and purpose, recognizing the genre of a document from its form provides information as to the document's purpose, which is otherwise difficult to assess.

1.3. Genres on the Web

The Web is an interesting setting for studying genres, because 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 Universal Resource Locator (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.

In many other cases, however, 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 is unlikely that there is a single Web community at all. Therefore, the resulting genre repertoire of a collection of Web pages will be the result of interactions among communities. In some cases, a genre may act as a type of boundary object [16], providing a common point of contact between different groups [6]. In others, this mixing may lead to confusion. For example, organizations have used the Web to publish information such as product brochures, annual reports, country, state, and city home pages, and government agency press releases. These organizations tend to use existing genres when putting information on the Web. A person happening to reach a document on one of their Web sites, however, 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. We propose that a major symptom of this genre confusion is the low success rate of Web searches.

1.4. How Automated Genre Recognition Can Help

Besides a rapid decrease in electronic storage costs, the last two decades of the previous century have witnessed a dramatic increase in computational power. As a result, improvements in computer based information technologies have made a reality many applications unthinkable just 20 years ago. Today's computers can recognize faces, human speech, and handwriting. Techniques have been developed to automatically classify documents according to topic sometimes surpassing humans in accuracy. These classification tools have been successfully using such technologies as machine learning algorithms and natural language processing [11].

The modern technology is ready to be advanced to the point where it can automatically recognize not only variation in content of textual information but also its form, in particular the genre of digital documents. Success of machine learning application to text classification tasks [11] allows us to believe in the possibility of recognizing digital genre with the accuracy comparable to that of a human being.

We believed that automated recognition of genre on the Web can dramatically leverage information seeking on the Web, thus addressing a real and growing problem. The users may be able to specify genres of interest or what tasks they are trying to accomplish. Based on this feedback, the search system may decide to promote or demote certain pages in the ranked order of retrieved documents. For example, it has been noted that most searchers are not interested in getting personal home pages [3], so the latter may be moved down the list by request. If someone is shopping on the Web, storefront e-commerce pages may be promoted while the pages containing educational material can be moved to the bottom of the list.

To use genre as an attribute in a Web search requires the development of computer algorithms able to recognize genre automatically. However, we quickly realized that recognizing all Web genre mentioned in the literature (e.g., the hundred or so identified by Crowston & Williams [4] would be infeasible. Fortunately, we believe that such precision may be unnecessary. Instead, it seems plausible that a large number of search tasks could be satisfied by documents of only a few groups of related genres, in which case distinguishing among documents in these groups would be almost as valuable as perfect recognition.

1.5. Our Contributions

In this paper, we present our first set of results, not published yet previously, from an on-going study aimed at facilitating World Wide Web searching through the automated recognition of genre. We present three closely related efforts. First, through our pilot exploratory study of web users, we identified what document genres they most/least frequently face in the process of searching and what document genres most/least frequently address their information needs. Second, based on our empirical findings, we have tentatively proposed collections of genres that are better suited for certain types of information

needs. This finding has been intuitively expected but not previously explored in the Web context.

The next section describes our study. The final "4. Proposed User Interface for Genre-based Searching " section presents a possible user interface for genre-based web navigation. This interface supports three major functionalities: 1) limiting search to specified genres 2) visualizing the tree-like hierarchy of genres discovered in the search results and 3) accepting user feedback on the relevancy of the specified genres.

2. Update on Web Genre: an Explorative User Study

In this section, we briefly describe a study we conducted to determine if average Web searchers understood the concept of document genre, to characterize the reasons why users searched for documents on the Web and to see which genres were more or less useful in satisfying their searches.

2.1. Objectives

We were interested in looking at the problems and situations that drove users to search the Web, thus this study was conducted in the context in which individuals had specific problems that they wanted to solve using the Web. We primarily desired to answer the following research question: *Does the notion of genre help Web searching?*

In addition, we addressed a number of sub-questions, two of which are touched upon in this paper: *1) What kinds of problems do users try to solve using the Web? 2) Is there any association between purpose of searching the Web and genres existent on the Web?*

2.2. Design

To answer these questions, we conducted a series of interviews with Web users. The interviews were performed by thirteen (13) graduate students in the fields of Information Technologies between March 1 and March 22, 2000.

Subjects were selected by the interviewers. The interviewers approached individuals in campus computer labs, public libraries, and workplace settings engaged in searching the Web. At the beginning of an interview, respondents were asked to describe the problems that they were using the Web to solve and to anticipate the genres of Web pages that would contain useful information for solving those problems.

Interviewers also collected the pages examined by respondents during the interview, collecting a total of 1234 Web pages altogether. Most of those pages were returned by search engines of respondents' choice or by following links from such pages. Respondents were asked for their assessment of the genre of the pages they found, the utility indicators they used, and whether those Web pages were relevant. This part of the interview could have multiple loops depending on the number of pages a respondent had looked at for information. At the end of interviews where respondents were asked for demographic information.

Responses were obtained from 184 individuals. Among the respondents, there were 87 males and 97 females, 91% had received at least some higher education, 52% had received graduate level education. 99.5% of respondents had used Web before and approximately 72% of them evaluated their proficiency of using the Web as above average. More than half of respondents worked or studied in an academic environment.

In the next step, the interviewers assigned a genre code from a content analysis scheme to each of the 1234 pages found by a respondent. The initial genre scheme was taken from [4] and user-articulated genres during interviews were added to it to form the final scheme. Each interviewer coded the data from his/her interviews, and data sets from all interviewers were compiled into one big set.

2.3. Findings

The first question we sought to answer regarded the *purposes for which users searched on the Web.* Table 1 presents the top most represented purposes in the sample. We found that Scholarly Research (incl. listings, full-text articles, pictures, and research news.), Shopping (incl. auctions, gifts, products, and real estate), Cultural Arts Activities, Health, News, Computing, Travel Planning, and Hobbies/Avocations were the most common purposes that respondents reported.

The second question we sought to answer was *whether there was a relation between the purpose of a respondent's search and the genre of the documents retrieved*. To answer this question, we first identified the genre of the retrieved documents. Altogether, 116 different genres were identified. Of this total, 72 were included in [4]

Table 1: Summary of purposes pursued in Web searching.

PURPOSE	PERCENTAGE	PURPOSE	PERCENTAGE
Scholarly Research	22.95%	Reference	3.28%
Shopping	12.57%	Countries	2.73%
Cultural Arts Activities	7.10%	Job-Hunting	2.73%
Health	7.10%	Educational Pursuits	2.19%
News	7.10%	Self-Help	2.19%
Computing	6.56%	Leisure Time/Recreation	1.64%
Travel Planning	6.56%	Playing	1.64%
Hobbies/Avocations	4.37%	Financial	1.09%
Curiosity	3.83%	Listening	0.55%
Reading	3.28%	Spiritual Enrichment	0.55%

classification and 44 (more than 1/3 of the total) were new ones articulated by respondents. None of the additional genres were added at the top level of [4] classification, except for a "Miscellaneous" group composed of genres that the researchers did not fit into any major groups in the traditional hierarchy (for example *calendar, contact, discussion-board/forum, schedule, weather page, and stock quotes).* Please see (http://www.ist.syr.edu/~roussinov/genre/appendix.html) the complete content analysis scheme.

To our surprise, we did not find a significant number of personal home pages in the sample of pages retrieved by respondents. We can explain this result by the fact that the Web search engines do not return many personal homepages, a finding very much in line with the observation made by Lawrence and Giles [10]. They found out that commercial search engines are typically more likely to index sites that have more links to them. Thus, automated recognition of personal home pages for the purpose of promoting or demoting them in the rank-ordered lists of retrieved documents does not seem to be necessary if the genre based system is built based on commercial search engines.

In order to determine whether document genres make sense to average Web users, we measured the *agreement between the genres assigned by interviewers and by the respondents*. Unfortunately, not all pages could be classified. In some cases, respondents gave answers such as "I don't know" or answers that did not have anything to do with genre. In others, the interviewers could not retrieve the Web pages at the time of their assessment. These uncoded pages were excluded from the sample for this particular analysis. Genre was identified for 1076 pages. In addition to coding the respondents' genre labels, the researchers also independently coded the same 1076 pages. The agreement between respondents and coders was 49.63%, which is higher than could be attributed to chance, but still low.

We noted that many of the disagreements were due to more specific use of terms by coders than by respondents. In other words, many disagreements were hierarchical rather than disagreements in kind. More generally, since some genres are similar to each other (e.g., *news bulletin* and *press release*) we would like to develop an agreement metrics that takes this similarity into consideration. In summary, we feel our data suggests that there is general agreement among Web genres perceived by different people.

However, the observation that the agreement is not perfect suggests that the notion of Web genre is indeed a little fuzzy, thus an interactive interface has to support some degree of flexibility (fuzziness).

Table 2 presents the most frequently reported purposes for searching the Web, along with the genres of the document that respondents found to be relevant to those tasks. We observed that

documents of see that particular genres were seen to be good for particular purposes, suggesting that the ability to search for documents by genre could be useful.

3. Web Genre to Facilitate Searching

As mentioned above, perfect automatic recognition of a large number of document genres does not seem to be feasible due to the limitations of the current state of the art of Artificial Intelligence and to the subjectivity of definitions of each genre. However, it seems plausible that a large number of search tasks could be satisfied by documents of only a few groups of related genres,

Table 2. Summary of genres associated with specific purposes.

STATED PURPOSE	GENRES THAT WERE FOUND MOST FREQUENTLY ASSOCIATED WITH THAT PURPOSE
Cultural Arts Activities	Announcements, organizational/business home pages, discography, search forms, topical home pages, search result lists, reports, articles
Scholarly Research	Table of contents, articles, topical home pages, essays, annotated hyperlink result lists, bibliographic records, books, tutorial pages, slides, press releases
Shopping	Product information, advertisements, organizational/business home pages, product lists, search result lists, table of contents
News	Organizational/business home pages, articles, search result lists, records, product information, topical home pages, newswire articles, stock quotes page
Health	FAQ, articles, search forms, discussion forms, product information, topical home pages, reports, self-help page, search result lists
Travel Planning	Organizational/business home pages, guidebooks, annotated hyperlink result lists, search forms, weather page
Hobbies/Avocations	Organizational/business home pages, graphic pages, search forms, search result lists, table of contents
Computing	Tutorial page, software downloading page, topical home pages, search result lists, project lists, publication lists, product reviews
Countries	Guidebooks, organizational/business home pages, map page, table of contents
Reading	Search result lists, summaries, chronicles, organizational/business home pages, tutorial page, vitae
Job-Hunting	Organizational/business home pages, search forms, prospectuses
Self-Help	FAQ, directories, instructions, link lists, table of contents
Leisure Time/Recreation	Sports records, link lists, topical home pages, reviews, newsletters
Educational Pursuits	Organizational/business home pages, table of contents, entry / referral page

in which case distinguishing among documents in these groups would be almost as valuable as perfect recognition. In short, even a good guess as to the purpose of a page may help those overloaded with information pick out more useful pages.

3.1. Objectives

In this section, we describe our attempts to identify 5-6 major groups of Web genre that satisfy the following two properties: 1) The relationship of each page to each of the groups can be established with the accuracy in the range of 80%-100% and 2) each group is likely to satisfy a limited subset of the observed search purposes. In the rest of the section, we describe how we arrived at such a set of groups through intuitive refinement. We realize that such a set may not be optimal. Combined with an appropriate user interface (next section), however, they may results in increased searching performance.

3.2. Procedure

To develop the set of document genres, we first listed possible indicators of genre that could be automatically recognized with an accuracy of 60-100%. Second, we associated those indicators with each genre related to the major purposes in the sample (listed in Table 2). At this point, these accuracy estimates are based solely on our intuition, since we have not yet begun to do any programming. Later, we may also involve more subtle indicators. For example, it is known that even simple text statistics such as frequency of long words or number of complex conjunctions have been found to be indicative of text genres [9]. Achievements in text authorship determination, going back to the early years of statistical stylistics [12], may also be relevant to automatic genre recognition. For example, it has been found that such simple statistics as relative pronoun counts can predict the formality or informality of a text [9].

3.3. Outcome

Finally, again following our intuition, we started to group genres into groups trying to facilitate automated recognition between groups as much as possible. After a number of refinements and rearrangements, we arrived at the groups listed in the Table 3. The table lists the five groups, the included genres and the possible indicators that may be used for automated distinguishing among those groups. These groups should be viewed as our initial proposal for implementation rather than a finished result. Unfortunately, at this stage of our on-going project, we can not give an exact numerical estimate of recognition errors. We assume that these groups will evolve as we implement and experiment with automated recognition techniques.

4. Proposed User Interface for Genre-based Searching

Based on the findings described in the previous sections and observations stated in the vast body of literature in the area of graphical interfaces for information access (see for example [8]), we have begun to design a novel graphical user interface that supports genre based Web searching. In addition to the search mechanism behind scene, the quality of the user interface can substantially affect user's experience with information access [8]. Our interface design aims at utilizing the advantage of genre while ensuring intuitive and easy-to-navigate page layout. To achieve these goals, the interface enables users to 1) limit searches to specified genres, 2) visualize the hierarchy of genres discovered in the search results, and 3) provide user feedback on the relevancy of the specified genres.

We have not yet implemented a system, but we can describe how it might be used. In the examples presented in Figure 1 and 2, a hypothetical user wanted to explore the financial situation of Coca-Cola company as reflected in the news. She entered the query "Coca-Cola." After the user submits an initial query, a rank ordered list of retrieved documents along with their brief summaries (snippets) is displayed on the right side in a way similar to traditional search engines, for example AltaVista (www.altavista.com). An example of the initial rank order of documents is shown on Figure 1.

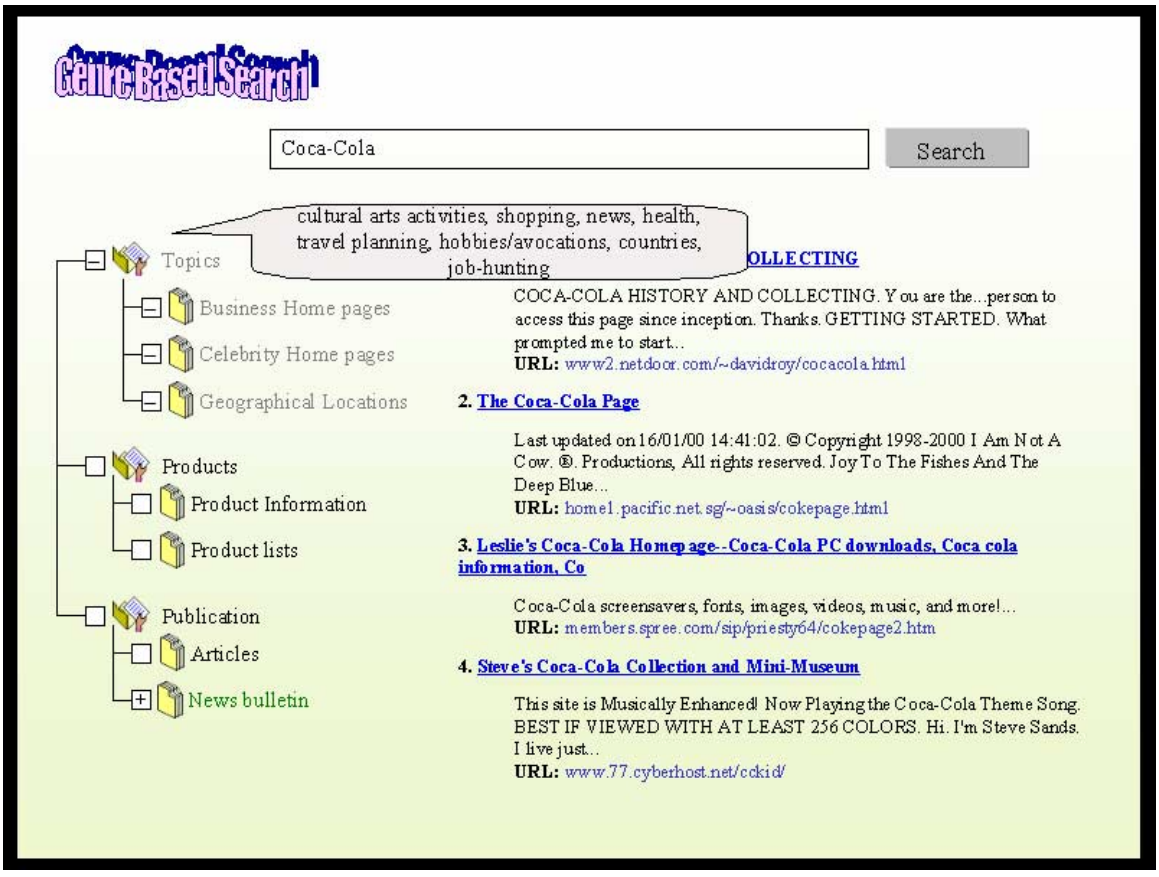

Figure 1. Tree-like visualization of genres found in search results.

Table 3. Recognizable groups of Web genres along with their indicators and the purposes they serve.

GROUP	MAJOR GENRES INCLUDED	MAJOR TASKS ADDRESSED	RECOGNITION INDICATORS
Topics	Home page (business, celebrities), geographical location, special topics	Cultural arts activities, shopping, news, health, travel planning, hobbies/avocations, countries, job-hunting	url consists only of a host name (f. e. juliaroberts.com) short in length plenty of graphics and ads large number of incoming links
Publications	Articles, publications, news bulletins	Scholarly research, news, financial	structure: hierarchy of sections longer than average not many graphics or ads url contains digits
Products	Product information, product lists, advertisements, ratings, product reviews, order forms, product lists	Shopping, news, computing	short in length prices or phone numbers url contains digits plenty of graphics and ads
Educational material	Glossary, course lists, instructional materials (guidebooks, instructions, manuals, problem sets, syllabus, tutorial page)	Educational pursuits	.edu domain education related lexicon (f. e. lecture, textbook, course, etc.) not many graphics or ads
FAQ	FAQ	Health, self-help	Special keywords inside urls, metadata and headings such as *FAQ,Q&A, questions and answers* structure (questions, followed by short answers) not many graphics or ads

In addition to simply retrieving documents, the set of genres detected in the documents matching the query is visualized in a hierarchical fashion (Figure 1). In the example, the system detected three major groups of document genres in the documents matching this query, *Topics*, *Products*, and *Publications*, while no *FAQ*s or *Educational Materials* were detected.

When the user places mouse over a genre a short description of that genre pops up. The user may mark the checkboxes of relevant genres with "+", irrelevant with "-" or leave some checkboxes blank refraining from an opinion. The hierarchical presentation allows the user to select several genres at once rather than having to inspect each one individually. Providing the feedback change the label's color—relevant genres are green, irrelevant gray.

In the example, shown in Figure 2, the user then marked *Topics* as irrelevant, *New Bulletins* as relevant and left other genres neutral. Please note that this interface is much less restraining than the one based on the notion of a folder, e.g., as supported by the NothernLight (www.northernlight.com/) search engine. A similar design for handling user feedback has been shown to be effective for navigation on the Web based on automated topical clustering and summarization [15].

Once the user provides feedback, the rank order of retrieved pages changes to promote documents of the relevant genre and demote irrelevant ones (Figure 2). In the example, the system reordered the search results as shown on the Figure 2. As a result of the feedback, only news Web pages showed up at the beginning of the list.

5. Conclusions, Limitations and Future Research

We reported on our ongoing study of using genre of Web pages to facilitate information exploration. Through an exploratory user study, we identified which genres most/least frequently meet searchers' information needs. Our initial results suggest that certain genres are better suited for certain types of needs. Based on these results, we have tentatively identified five (5) major groups of genres that can be used in an interactive search tool that would allow genre-based navigation. We suggested that each group can be automatically recognized by a computer algorithm based on the indicators that we also suggested. Also each group has a better chance of satisfying particular types of information needs, thus the user may narrow down his/her search. A second conclusion from our empirical works is that the users do not always agree on a genre. We believe that an interface supporting fuzzy genre definitions will therefore be more usable. We have proposed an interface that allows genre-based navigation through three major functionalities: 1) limiting search to specified genres 2) visualizing the hierarchy of genres discovered in the search results and 3) accepting user feedback on the relevancy of the specified genres.

Our study is still in its early stages and the preliminary results presented here have numerous limitations that we will address in future studies. First, our sample is yet too small to use as training set for a machine learning algorithm and achieve reasonable accuracy. For this and other reasons, we have not yet attempted to implement any recognition algorithms.

Second, we noticed that some interviewers tried to force the genres of Web pages into the genre scheme we adopted from [4] even though they did not match very well. This over-reliance on the prior categorization scheme may have introduced a potential bias for our user-based approach. To address this, it will be necessary to use an inductive scheme in the future.

However, even with these limitations, our results seem to justify further development of genre-based navigational tools for the Web. Considering the ever-increasing torrent of information, incorporating a dimension of genre into the search process may be a rescue.

As a next step, we are planning to collect a large sample of user-selected and evaluated web pages, manually code the indicators of genre in the obtained pages using an inductive content analysis procedure and test machine learning algorithms for their accuracy. As a parallel activity, we are planning to run user experiments with data collections where genre information is *a priori* known to test how knowing genre may help to narrow down the search. Finally, we are hoping to unite those two activities into a next generation Internet search engine.

References

[1] Bowman, C.M. The Harvest information discovery and access system. *Proceedings of the Second International World Wide Web Conference '94*, Chicago, IL.

[2] Campbell, K. K. and Jamieson, K. H. (Eds.). *Form and Genre: Shaping Rhetorical Action*. Falls Church, VA: Speech Communication Association.

[3] Chen, H., Schuffels, C., & Orwig, R. Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation,* 7(1), pp. 88-102.

[4] Crowston, K. and Williams, M. Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society*, 16(3).

[5] DeBra, P. & Post, R. Information retrieval in the World-Wide Web: Making client-based searching feasible. *Proceedings of the First International World Wide Web Conference '94*, Geneva, Switzerland.

[6] Freedman, A. and Medway, P. Locating genre studies: Antecedents and prospects. In: A. Freedman and P. Medway (Eds.), *Genre and the New Rhetoric*, pp. 1-22. London: Taylor and Francis.

[7] Harrell, J. and Linkugel, W. A. On rhetorical genre: An organizing perspective. *Philosophy and Rhetoric*, 11, pp. 262-281.

[8] Hearst, M.A. User Interfaces and Visualization**,** in *Modern Information Retrieval*, edited by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley Publishing Company.

[9] Karlgren, J. and Cutting, D. Recognizing text genres with simple metrics using discriminant analysis. In: *Proceedings of COLING 94*, Kyoto.

[10] Lawrence, S and Giles, L. Accessibility of Information on the Web, *Nature*, 400, pp. 107--109.

[11] Lewis, D. and Sparck Jones K. Natural language processing for information retrieval. *Communications of the ACM*, 39(1), pp. 92-101.

[12] Mendenhall, T.C. The characteristic curves of composition. *Science*, 9, pp. 237-49.

[13] Miller, C. R. Genre as social action. *Quarterly Journal of Speech*, 70, pp. 151-167.

[14] Orlikowski, W. J. and Yates, J. Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly*, 33, pp. 541-574.

[15] Roussinov, D. and McQuaid, M. Information navigation by clustering and summarizing query results. In: *Proceedings of Hawaii International Conference on System Sciences (HICSS-33)*. January 4-7, Maui.

[16] Star, S. L. and Griesemer, J. R. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19, pp. 387-420.

[17] Swales, J. M. *Genre Analysis: English in Academic and Research Settings*. New York: Cambridge University Press.