Syracuse University
SURFACE

School of Information Studies: Faculty Scholarship | School of Information Studies (iSchool)

12-1-2000

Identifying Document Genre to Improve Web Search Effectiveness. The Bulletin of the American Society for Information Science and Technology

Barbara H. Kwasnik
Syracuse University, bkwasnik@syr.edu

K. Crowston
Syracuse University, kcrowston@syr.edu

Mike Nilan
Syracuse University

D. Roussinov

Follow this and additional works at: https://surface.syr.edu/istpub

 Part of the Library and Information Science Commons

Recommended Citation

Kwasnik, Barbara H.; Crowston, K.; Nilan, Mike; and Roussinov, D., "Identifying Document Genre to Improve Web Search Effectiveness. The Bulletin of the American Society for Information Science and Technology" (2000). *School of Information Studies: Faculty Scholarship*. 133.
https://surface.syr.edu/istpub/133

This Article is brought to you for free and open access by the School of Information Studies (iSchool) at SURFACE. It has been accepted for inclusion in School of Information Studies: Faculty Scholarship by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

Identifying Document Genre to Improve Web Search Effectiveness

by Barbara H. Kwasnik, Kevin Crowston, Michael Nilan and Dmitri Roussinov

The authors are affiliated with the School of Information Studies at Syracuse University, Syracuse, NY 13244-4100. They can be reached at the following e-mail addresses: Barbara H. Kwasnik, bkwasnik@syr.edu; Kevin Crowston, crowston@syr.edu; Michael Nilan, mnilan@syr.edu; and Dmitri Roussinov, droussin@syr.edu

Communication technologies have enabled people to create millions of electronic messages and Websites representing a variety of interests. Personal home pages and business sites have in recent years been accruing at a rate of hundreds of thousands a day. Our productivity in generating information has exceeded our ability to process it, and the dream of creating an information-rich society has become a nightmare of information overload. The advantages of huge stores of information are often outweighed by the difficulties of accessing them.

Stories of searching dysfunction abound. For example, finding the answer to the question "How long does it take to get by train from Copenhagen to Oslo?" seems possible through composing a query "copenhagen AND oslo AND train." A search on Alta Vista results in about 900 matching Web pages, only one of which contains the answer. Most users tolerate exploring only about the first 20-30 pages in the ranked lists before giving up.

It is in this environment that our research team has set out to improve Web searching by developing a method for automatic genre identification. We believe that the ability to identify a document's genre may improve the precision and utility of searches by making it possible to find documents that not only match the user's search terms, but also include (or exclude) documents of a given genre or genre cluster.

The Notion of Genre

The notion of genre is not new. 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, but we build on the one proposed by W. J. Orlikowski and J. Yates. They define genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form." 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. Most genres, however, imply a combination of purpose and form, such as a newsletter, which communicates "the news of the day" by including multiple short articles and is also distributed periodically to subscribers or members of an organization. Because genre implies both form and purpose, recognizing the form of a document provides information as to the document's purpose, which is otherwise difficult to assess.

Genres are useful because they make communications more easily recognizable and understandable by recipients, but efforts to describe genre face many challenges. For one thing, many traditional classification systems have awkwardly tried to describe the genre of a document using schemes that are basically designed to describe topic rather than form. In other words, traditional systems have asked, "What is this document about?" rather than "What type of document is it?" As new media have been incorporated into both our print and electronic collections, the need for a way to describe form has acquired new urgency.

Some movement in that direction has already been made. The Dublin Core of Metadata Elements, for instance, makes provision for an element to describe "document type." The plan in this metadata effort is to use a controlled vocabulary of terms from which to choose. The "type" element, however, requires further research and definition, and we anticipate that our work will contribute to that effort as well.

Genres and Classification

We proceed, therefore, from an assumption that documents can be usefully classified by their genre.

Classification allows people to talk about basically dissimilar things using a common label. For example, while each instance of a letter is different from any other instance, and a form letter is different from a personal letter, it is often helpful to use the term *letter* to cover the entire set of letters, despite distinctions. The aim of classifying is to identify salient criteria for the association of similar items and the critical distinctions among them. Thus in identifying and classifying genres we must take into account many factors. Among these is the nature of the objects themselves, the cultural context, the function and uses of the classification, as well as required granularity and scale. Put another way, objects, such as genre types, can be viewed from many angles and be a part of many intertwined classification schemes.

Genres on the Web

The Web is an interesting setting for studying genres, because there are many communities meeting on the Web, bringing with them experiences with a variety of 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. 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 what S. L. Star and J.R. Griesemer call a type of boundary object, providing a common point of contact between different groups. 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 Websites, 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 this genre confusion may, in part, account for the low success rate of Web searches.

How Artificial Intelligence Can Help

The last two decades have witnessed a dramatic increase in computational power. As a result, improvements in computer-based information technologies have led to the development of many applications that were unthinkable just a short time ago. Today's computers can recognize faces, human speech and handwriting. Techniques have been developed to automatically classify documents with respect to their topic. Such automated classification tools have been successfully using techniques such as machine learning algorithms and natural language processing. Increasingly, these techniques have also been applied to the visualization of large volumes of scientific data and texts.

We think that this technology is ready to be advanced to the point where it can automatically recognize not only variation in the content of textual information but also its form, in particular the genre of digital documents. The success of applying machine-learning techniques to text-classification tasks encourages us to think that it can also be applied to the identification of genre and that this application will markedly improve searching in very large and dynamic environments, such as the Web.

The Research Plan

Our work addresses four questions:

1. Does using information about a document's genre improve the effectiveness of Web searches?
2. What document genres are used and sought by searchers on the Web?
3. Can document genre be automatically defined?
4. How can genre information be used in searching and how can the interface to the new functionality be implemented most effectively?

The following is a brief summary of the major components of our multi-facetted approach:

Question 1: Identifying Genres Used on the Web

There are several sources for identifying genres on the Web. As a starting point we build on the work of K. Crowston and M. Williams, who developed a sample of Web documents and content analyzed them into identifiable genres. Since their sample was taken several years ago, and so may not fully reflect the current uses of the Web, a new pool of documents will be generated by random sampling techniques, including ones based on real users' searches. Classification of these documents will expand the range of genres identified by Crowston and Williams as well as verify the genres that they have already identified. We anticipate classifying about 2000 pages in order to create a large pool of classified pages.

Then, we will conduct a descriptive study of Web users' actual search behaviors. Working with subjects with real searches, we will note the purpose behind the user's search and ask the user to identify "indicators" or characteristics of Web pages. We will examine the actual language employed by users to refer to various genres. Users may recognize a genre and its traditional (or emergent) connotation and yet refer to it by a label different from more traditional sources. In order for our efforts to be functional in practice, we will document the differences in linguistic labels associated with specific genres.

Finally, the situation or problem that leads a user to search

the Web is potentially indicative of differences in patterns of genre use among different Web communities. A preliminary study of over 1,000 Web pages has yielded encouraging results by adding to our repertoire of known Web genres, and by identifying how users talk about these genres. Thus, the descriptive study of Web users intends to identify user-based genres, content-based relevance indicators of those genres, and the language users employ to label the genres. The associations among these three elements will constitute usability criteria for evaluating subsequent searches conducted with genre-based search algorithms. Past research has shown that situational constraints affect both the language employed by users to refer to resources and their judgments of the utility of retrieved resources.

Question 2: Classifying Web Genres

Having collected a large number of examples, we will need to create a typology of genres on the Web using facet analysis. Facet analysis identifies multiple fundamental dimensions along which objects, such as genre types, can be described. Each facet is articulated following its own logic. This approach allows for the development of description and clustering using a number of fundamental dimensions, rather than just one, as in traditional classifications. The results of this process will yield a classification that is flexible, expressive and hospitable to new genre types and genre combinations. It will also allow us to view genre types at a variety of conceptual levels, from general and inclusive, to very specific.

The resulting range of patterns will then be compared to the more traditional sources of genre types for overlaps in structure and coverage. This step will provide us with a robust and comprehensive set of specifications for subsequently being able to identify document genres automatically. Our intent is to generate a classification that reflects not only currently identifiable genres but that will also flexibly accommodate identification of future genres. The results of this work include a range of situated genre types that are representative of actual Web use, their associated content-based relevance indicators, a classification scheme for the genre types and a database of Web pages for training and evaluation of the genre classification software, described next.

Question 3: Software to Automatically Identify Genre

Once we identify user-defined genres on the Web, enriched by contextual information relating to use, we are still faced with the problem of applying the results in a retrieval situation so vast and dynamic it is simply impractical to manually index each document. For this reason, our team will pursue auto-

matic genre identification in parallel with manual document and user-based identification. We will develop computer software to automatically recognize the genre of documents by exploiting observed regularities of substance and form. This work will follow two complementary approaches, heuristic and machine learning.

The Heuristic Approach. Based on the findings of the Web-genre characterization activity, we will build a set of heuristic rules to identify popular genres. For example, a FAQ document usually contains the words "FAQ" or "Frequently asked questions" and is formatted as a series of questions and answers. Many home pages have "~" (tilde) in their Web address and the words "home page" in their titles. Even simple text statistics such as frequency of long words or number of complex conjunctions may be indicative of text genres.

Machine Learning Approach. A machine learning classifier is able to learn from a set of input-output pairs. The inputs are objects (for example, Web pages) described by their features (for example, structure, length, URLs and language). The outputs are memberships (true/false) in a particular class (for example, a class of documents classified as "letters"). Having learned the given input-output pairs, the system can classify inputs that it has never seen before, sometimes surpassing humans in accuracy of predicting the outputs. We will use half of the pool of pages manually classified in the first activity to train the classifiers.

The manual and automatic methods of identifying and defining genre will be conducted iteratively, each technique contributing to the other.

Question 4: Genre-Based Navigation and Integrated Testing

To test the utility of genre for improving Web search effectiveness, we will integrate automated genre recognition with a search engine. This activity will involve research on how genre information might be used in searching and on the interface to the new functionality. The most straightforward approach would be to allow users to select the genres in which they are interested. The search engine could then filter out documents with irrelevant genres.

Finally, we plan to undertake a formal evaluation of the utility and usefulness of an interactive Web search engine augmented with a genre identification module. We will run a series of experiments to compare the performance of existing search engines with the genre-augmented search system.

Why We Want to Do This

We think that our proposed approach holds some promise for improving Web search-engine performance, thus addressing a real and growing problem. Our description of genres should be useful to both professional and amateur Web designers seeking to make their content more readily understandable and accessible to users. Researchers working in closely related areas, such as Web searching or interactive information retrieval, will benefit from the metrics that we will introduce. The approach we propose, although designed and tested primarily for Internet users, can be later adapted for searching, categorizing and visualizing company intranets, e-mail repositories and legacy documentation. In addition, if we are successful, our conceptual and methodological approach will represent a dramatic improvement in studying information, information use and communication phenomena on the Web. In solving the pragmatic problem of imprecise Web searching, we will also improve our collective understanding of how document genres are being used and recreated in a novel medium, thus illuminating the underlying social dimension of communication behavior.

Further Reading

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

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

Dillon, A. & Gushrowski, B. (2000). Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science*, 5, 202-205.

Freedman, A. & Medway, P. (Eds.) (1994), *Genre and the new rhetoric*. London: Taylor and Francis.

Harrell, J. & Linkugel, W. A. (1978). *On rhetorical genre: An organizing perspective*. Philosophy and Rhetoric, 11, 262-281.

Karlgren, J. & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In: *COLING 94: Proceedings of the 15th International Conference on Computational Linguistics*, August 5-9, 1994, Kyoto. N.p: ICCL, 1071-1075.

Kwasnik, B. H. (1999). The role of classification in knowledge representation and discovery. *Library Trends*, 48(1), 22-47.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, 70, 151-167.

Orlikowski, W. J. & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly*, 33, 541-574.

Roussinov, D. & McQuaid, M. (2000). Information navigation by clustering and summarizing query results. [CD-ROM]. In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS-33). January 4-7, 2000, Maui, Hawaii*. Los Alamitos, CA: IEEE Computer Society.

Star, S. L. & Griesemer, J. R. (1989). Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19, 387-420.

Yoon, K., & Nilan, M.S. (1999). Toward a reconceptualization of information seeking research: Focus on the exchange of meaning. *Information Processing & Management*, 35, 871-890.