Briefly Noted

Genres on the Web: Computational Models and Empirical Studies

Alexander Mehler,* Serge Sharoff,‡ and Marina Santini¶ (editors)
(*Goethe-Univerität Frankfurt-am-Main, ‡University of Leeds, and ¶Stockholm)

Berlin: Springer (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 42), 2010, xiv+362 pp; hardbound, ISBN 978-90-481-9177-2, $189.00

What are the typical genres on the Web? How can they be distinguished? What is the current application of genre theory to Web technologies? These are the sort of issues that would lead you to *Genres on the Web*, a collection of essays with a no-nonsense title, edited by Alexander Mehler, Serge Sharoff, and Marina Santini. If you understand categorization, but need to know something about the Web and text genres, this book is a pretty good and accessible starting point.

The individual essay authors are generally well-respected experts in the field and the 16 individual essays are all solid pieces of work in their own right. The collection is well-structured with an introduction by the volume editors and four sections on, respectively, "Identifying the Sources of Web Genres," "Automatic Web Genre Identification," "Structure-Oriented Models of Web Genres," and "Case Studies of Web Genres."

The bulk of the material is in the genre identification section, and could serve as the basis for a seminar in text classification more generally, as it illustrates many of the major issues and themes in text categorization. The target reader is assumed to be familiar with much of the underlying technology; for example, several authors (Santini, Chapter 5; Kim and Ross, Chapter 6; Sharoff, Chapter 7; Stein et al., Chapter 8; Braslavski, Chapter 9) talk about support vector machines, but none define or describe them. Of these authors, only Braslavski discusses alternatives, and then only in passing. This should not prove a major barrier. Other sections deal effectively with issues such as collecting and normalizing Web samples for genre research, the relationships between and among genres, and the evolution of "style" on the Web. Of particular note is the extensive collection of figures, diagrams, and tables illustrating the results of the various analyses, which help make this a particularly accessible book to the lay reader.

There are two weaknesses in the collection. The first is the (lack of) discussion of the theory of genre itself; there is no clear definition of what exactly constitutes a "Web genre" aside from the simple enumeration of categories listed in benchmark corpora. Of course, this lack of clarity in the collection is a consequence of an equivalent lack of clarity among scholars of genre, as pointed out by the paper by Rosso and Haas, and this weakness should not be a stumbling block to most of the readership of *Computational Linguistics*. (If anything, it shows where more work on the theory of genres is needed.)

The second weakness is the relative lack of discussion of genre in non-text documents. Aside from one essay (Paolillo, Warren, and Kunz) on Flash video, there is almost no attention paid to the multimedia aspects of Web publishing. This may be a more serious problem given that many of the papers describe their technology and applications in terms of improvements to Web search technologies, and text Web search is already so much more advanced than music or video search that there may be more and better applications (lower-hanging fruit) available in other areas.

Despite these relatively minor weaknesses, the book remains a good reference for the current state of the art in the study of genre on the Web, and will be of interest to anyone trying to sort out a computational notion of "genre," on or off the Web.—*Patrick Juola, Duquesne University*