Categorizing Web Viewership Using Statistical Models of Web Navigation and Text Classification

Alan L. Montgomery
Associate Professor, Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15241-3890.
E-mail: alan.montgomery@cmu.edu, Phone: 412-268-4562, Fax: 412-268-7357

Brett R. Gordon
Masters Student, Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15241-3890.
E-mail: brgordon@andrew.cmu.edu.

Abstract:

Clickstream data refers to the sequence of World Wide Web (WWW) pages accessed by a user, and provides a rich resource for understanding information search by individual users. However, this data is also unstructured, both in the lack of knowledge about why a user is browsing and the lack of information about a category. In this paper we propose a technique for classifying the content viewed by a web user. We merge two sources of information to classify page viewings. The first is based upon a Markov model of user browsing behavior. The second is based upon the text that occurs on the page. Previous researchers have primarily focused on approaching these problems separately. When these approaches are applied separately they generally result in fair to good accuracy. We show that when these information sources are combined together we have a high degree of accuracy in predicting the category of a web page. Our technique has applications both in improving search engines, web design, and understanding browsing behavior. Methodologically, this research shows how to extract value from clickstream data and textual information. Both of these sources of data are underutilized in marketing research today.

Keywords: Clickstream Data, Hierarchical Bayesian Models, Markov Process, Text Classification, Internet Marketing