Implications of the Recursive Representation Problem for Automatic Concept Identifcation in On-ine Governmental Information[3] ¤

Miles Efron, Gary Marchionini, and Julinang Zhiang
School of Information and Library Science
CB#3360, 100 Manning Hall
University of North Carolina
Chapel Hill, NC 27599-3360
*f*efrom, marchg@ils.unc.edu, junliang@email.unc.edu
August 18, 2003

Abstract

This paper describes ongoing research into the application of unsupervised learning techniques for improving access to governmental information on the Web. Under the auspices of the GovStat Project (http://www.ils.unc.edu/govstat), our goal is to identify a small number of semantically valid and mutually exclusive "concepts" that adequately span the intellectual domain of a web site. While this is a classic instance of the clustering problem [14] the task is complicated by the dual-representational nature of term-document relationships. Since documents are defined in term-space and vice versa, we may approach this as a document-or term-clustering problem. The current study explores the implications of pursuing both term- and document-centered representations. Based on initial work, we argue for
a document clustering-based approach. Describing completed research, we suggest that term clustering yields semantically valid categories, but that these categories are not suitably broad. To improve the coverage of the clustering, we describe a process based on document clustering.

Please see the PDF Document in the packet for the full research paper.

[3] ¤This research was supported by NSF EIA grant 0131824.

Discussion Points

31. What does the author mean by classification?

32. Does the research involve creation or implementation of a classification scheme?

33. How does the researcher use classification to improve the automated approach?

34. How do these methods compare to current human-generated approaches to classification?

35. How does the reported research expand our understanding of classification?

36. Does the research suggest an improvement over human-generated classification?

37. What do you think are the most important lessons learned in this research?

38. What do you think are the best practices reported in this research?

39. What would you recommend to the researcher as the next step in this approach?

40. Is there other related research that you would recommend the researcher become acquainted with?