VOLUME 10 (2006): ISSUE 1. DISCUSSION PAPER 3

Discussion of "Operationalising "Websites": lexically, semantically or topologically?"

Gaston Heimeriks

NIWI, Royal Netherlands Academy of Arts and Sciences
E-mail: **gheimeriks@gmail.com**
http://hcs.science.uva.nl/usr/heimeriks/home.html

The core problem that is identified in this paper is that no effective empirical operationalisation for the concept "Website" exists up to now. Consequently, progress in the research areas of webometrics is hindered. The authors point out that, although the notion of a Web-page can be operationalised in a straightforward way as the resource corresponding to a particular URL, the same is not true of "Website". There is no consensus as to what a "Website" is, although in practice the term is frequently used to suggest a coherent assemblage of Web-pages that is intended to be considered as a unit by the reader. Furthermore, there is a need to understand the effect that different pragmatic analytic choices may have on the results of such Web-graph analyses in order to aid interpretation of their results.

The scale and complexity of the Web-page digraph is typically managed by aggregating together or clustering individual Web-pages in order to form "Websites". It is the properties of these Websites which then become the focus of research. The most popular Web-page clustering technique is "lexical" and uses the URL syntax in order to assign Web-pages to "Websites". Semantic clustering, that is clustering Web-pages according to the similarity of their content has also been proposed. In this paper a third approach to Web-page clustering is suggested, which is based on the topological properties of the Web-page within the Web-page digraph. The technique and the results of an experiment to compare the use of URL-lexically and topologically determined Websites in two sub-domains are discussed. A benefit of this approach is that it avoids the administrative bias inherent in the URL-lexical approach. In consequence the topological approach is unaffected by the size of the Web host servers that are included in the Web-page digraph.

This explorative topological operationalisation provides a promising methodology. In particular such analyses could be useful for Websites that comprise of Web-pages from more than one host servers. Neither of the approaches demonstrated to operationalise "Website" (URL-lexical or topological) have exploited the (non-link) content of Web-page in order to aggregate Web-pages. Furthermore, as the current study focuses on the patterns within a websites, it would be interesting to study hyperlink network structures between topologically defined websites in order to gain more insight in the underlying patterns of hyperlinking.

MAIN PAPER	DISCUSSION