

# Topic Hierarchy Construction for Web Site Summarization

Yongzheng Zhang

Faculty of Computer Science

Dalhousie University

6050 University Ave.

Halifax, NS, Canada B3H 1W5

*yongzhen@cs.dal.ca*

## **Abstract**

Web document summarization, which identifies the essential contents of source documents, has gained much attention in recent years. In our previous work, we proposed a content-based system to generate a concise Web site summary by means of key phrase and key sentence extraction. Experimental evaluation shows that the automatically generated summaries can convey the same information as human-authored summaries do. However, as the size and diversity of Web sites continue to grow, straightforward summarization of an entire Web site may lead to an incoherent summary or a summary heavily biased to a subset of main topics covered in the target Web site. In this paper, we propose a new approach to summarization of Web sites with diverse topics and heterogeneous contents. The system is focused on creating a topic hierarchy which effectively organizes and visualizes the Web site contents. First a proper link hierarchy is constructed by capturing and refining the raw hyperlink structure of a given Web site. Second, documents in the link hierarchy are clustered into topic groups in a top-down manner. Third, topic groups are summarized and labeled to form a topic hierarchy in a bottom-up manner. We aim to apply extrinsic evaluation to measure the usefulness and effectiveness of the proposed approach in terms of how well the hierarchical summaries can help Web users understand the main topics and essential contents of Web sites.