

# Department of Computer Science and Software Engineering

## Master Thesis Defense

Speaker:	Tahira Hasan
Supervisor:	Drs. Mudur & Shiri
Examining Committee:	Drs. Ormandjieva, Radhakrishnan and Dr. Harutyunyan (Chair)
Title:	Finding Usage Patterns from Generalized Weblog Data
Date:	Tuesday March 31, 2009
Time:	2:00 pm
Place:	EV3.101

## ABSTRACT

Buried in the enormous, heterogeneous and distributed information, contained in the web server access logs, is knowledge with great potential value. As websites continue to grow in size and complexity, web usage mining systems face two significant challenges - accuracy and scalability. This thesis develops a web data generalization technique and incorporates it into a web usage mining framework in an attempt to exploit this information-rich source of data for effective and efficient pattern discovery. Given a concept hierarchy on the web pages, generalization replaces actual page-clicks with their general concepts. Existing methods do this by taking a level-based cut through the concept hierarchy. This adversely affects the quality of mined patterns since, depending on the depth of the chosen level, either significant pages of user interests get coalesced, or many insignificant concepts are retained. We present a usage driven concept ascension algorithm, which only preserves significant items, possibly at different levels in the hierarchy. Concept usage is estimated using a small stratified sample of the large weblog data. A usage threshold is then used to define the nodes to be pruned in the hierarchy for generalization. Our experiments on large real weblog data demonstrate improved performance in terms of quality and computation time.