CLE SEMINAR SERIES-III

Topic: Latin-Nastalique Script Classification System
Presenter: Mr. Muhammad Usman Ghani
Presentation Date: 15th July, 2014
Venue: Seminar Hall, KICS

Abstract:

In Urdu books and magazines, Latin script is also used for terminology illustration or other purposes. Therefore, script detection system has been developed which separates Latin and Nastalique script so that these can be recognized. Distinguishing features between both scripts have been determined. These features has been used to train a C4.5 algorithm based decision tree that classifies connected components (CCs) into Latin and Nastalique script CCs. Heuristics based neighboring rules has also been applied to further enhance the script classification process. Latin and Nastalique runs are then marked. The Nastalique script runs are recognized through Urdu OCR and Latin script runs are recognized through Tesseract OCR. System is trained on 983 pages and tested on 239 pages. Overall 99.61% script classification accuracy is achieved.