POST-PROCESSING OF LDA FOR CLUSTER QUALITY ANALYSIS

By

Matthew Fyffe

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Approved:

Sibel Adali, Thesis Adviser

Boleslaw Szymanski, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

April 2009
(For Graduation May 2009)

ABSTRACT

Latent Dirichlet Allocation (LDA) is a formal generative model of documents, breaking data into two tiers, documents comprised of keywords and topics comprised of distributions of keywords. LDA is implemented in most scientific papers employing information retrieval techniques, with emphasis being placed on modifying LDA or its inputs in order to improve its output.

While methods have been implemented to improve LDA's results, nothing has been done to analyze LDA's output to compare the results qualitatively. LDA clusters the data into topics but does not shed light on which topics are stronger than others. We propose a metric that executes multiple runs of LDA and compares the results to find the strongest topics in the data set.

Through simulations on both fake and real world data, we have shown that our proposed metric offers a reliable quality score for the topics.