SAICSIT '11, October 3–5, 2011, Cape Town, South Africa

.

Using N-grams to Identify Mathematical Topics in MXit Lingo

Laurie L Butgereit: Meraka Institute, Pretoria,RSA also
Nelson Mandela Metropolitan University, Port Elizabeth, RSA
lbutgereit@meraka.org.za

Reinhardt A Botha: Nelson Mandela Metropolitan University, Port Elizabeth, RSA
ReinhardtA.Botha@nmmu.ac.za

ABSTRACT

N-grams are used to quantify the similarity between two documents or the similarity between two collections of words. This paper shows how N-grams of length 3 and N-gramsof length 4 both coupled with text pre-processing (including stop word removal and stemming according to MXit spelling conventions) can be used to categorize very short mathematical conversations conducted in MXit lingo into broad mathematical groups suchas algebra, geometry, trigonometry, and calculus. MXit lingo is an abbreviated form of written English which children, teenagers and young adults utilise when communicating using the popular MXit chat mechanism over cell phones. Conversations from the “Dr Math” project were used for this analysis. “Dr Math” is a mathematics tutoring service which links primary and secondary school pupils to tutors from local universities. The tutors assist the pupils with their mathematics homework.