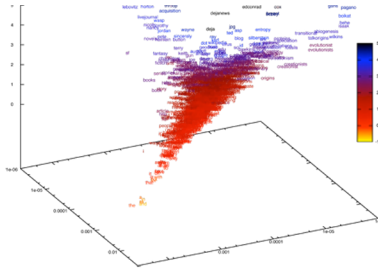


NORTHWESTERN INSTITUTE ON COMPLEX SYSTEMS PRESENTS

Wednesdays

@NICO

Language and Social Behavior in Usenet Groups



Eduardo G. Altmann, Northwestern Institute on Complex Systems, Northwestern University
Wednesday, October 22, 2008
12:00 – 1:00 PM
Technological Institute, 2145 Sheridan Road, Rm. M416 (Please note different location)!

Before the WWW, blogs, and IMs there were Usenet groups. This internet distributed discussion system has been used as a collective world-wide communication for the past three decades, building a detailed database of the interaction between millions of users. In this talk I will discuss how questions from linguistics and social behavior can be studied using Usenet groups. I start with a general characterization of the groups (e.g., the distribution of posts per user and posts per thread have heavy tail but the lifetime of users decays exponentially). I then discuss how the time dependent frequency of usage of specific words can be used to quantify the popularity of words, e.g., internet slangs, products, persons, or events. In the main application discussed in this seminar, I will take advantage of the amount of data available (~15 years and ~100,000,000 words in each group) to introduce a statistical characterization of words that goes beyond the frequency of usage. Based on the distance between successive occurrence of words, I will show that different parts of speech (in the same frequency range) can have different statistical properties: while function words follow approximately a Poisson process, content words consistently diverge from a Poisson process for both short and long distances. This motivates the definition of the area A between the measured and the Poisson distributions as a characteristic of the word usage in each group. Words with large A provide a good characterization of the discussion topics of the group. In terms of the different parts of speech, we find the following order for decaying A : proper nouns, common nouns, adjectives, verbs, and prepositions. This suggests a connection between the A -score and the semantic content of the words.

NICO Coffee Hour will follow for questions, networking, and collaboration.

<http://www.northwestern.edu/nico/>



NORTHWESTERN
UNIVERSITY