Linguistic and social patterns within online discussion groups

Robert Warren
School of Computer Science, University of Waterloo
rhwarren@uwaterloo.ca

David Evans
Faculty of Computer Science, Dalhousie University
dfe@cs.dal.ca

Abstract

In this research we analyzed the contents of several hundred on-line discussion groups from both a content and a social network analysis perspective. By studying the clusters of individuals within each discussion thread, we seek to determine whether cohesion in terms of group composition and writing style differs from one group to another. Previous studies of this type have not considered how behavior changes over time; we therefore identify the stability of the actors and of inter-actor relationships. Our methodology requires minimal meta-data from the discussion group infrastructure and combines this with syntactic and readability measures to form a rich characterization of the groups' social networks.

Introduction

We know from previous research on readability measures and social network analysis that communications networks reflect the behavior of people. In this work, we wished to review whether we could mix the two into a project to reflect on the patterns within online discussion groups. Hence we review here the results of our analysis of USENET newsgroup data based on the structural of online discussions and the overall complexity of the underlying text.

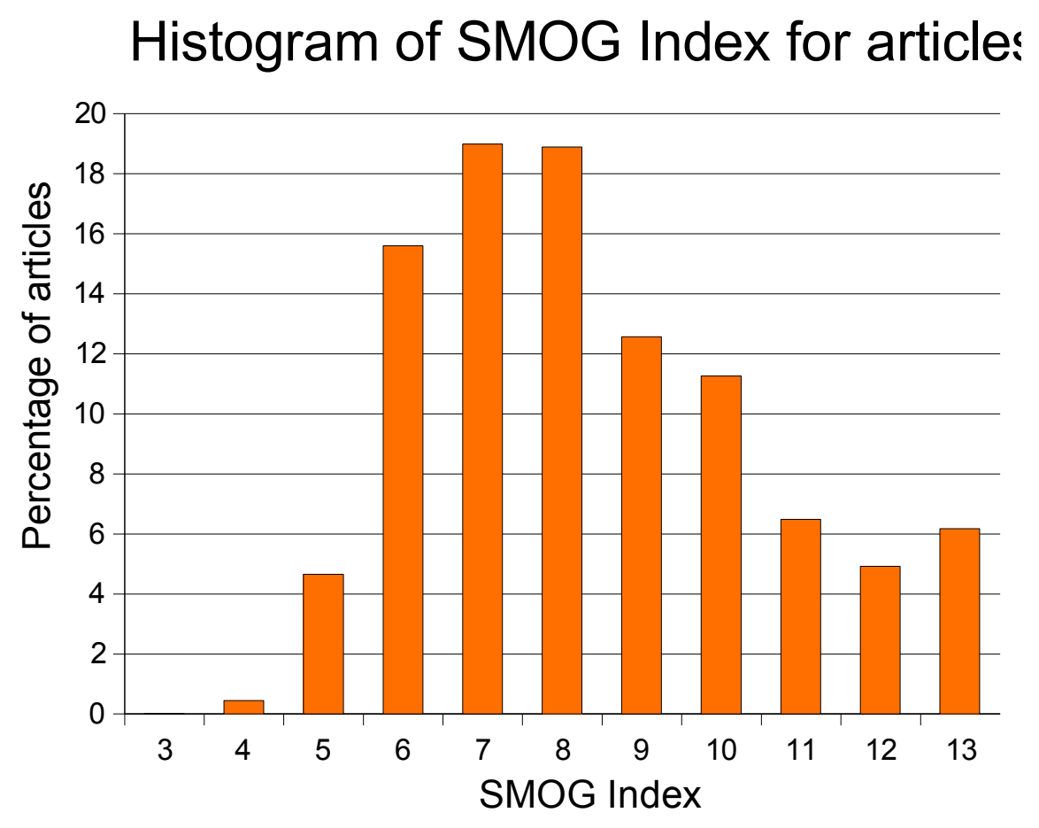
Dataset description

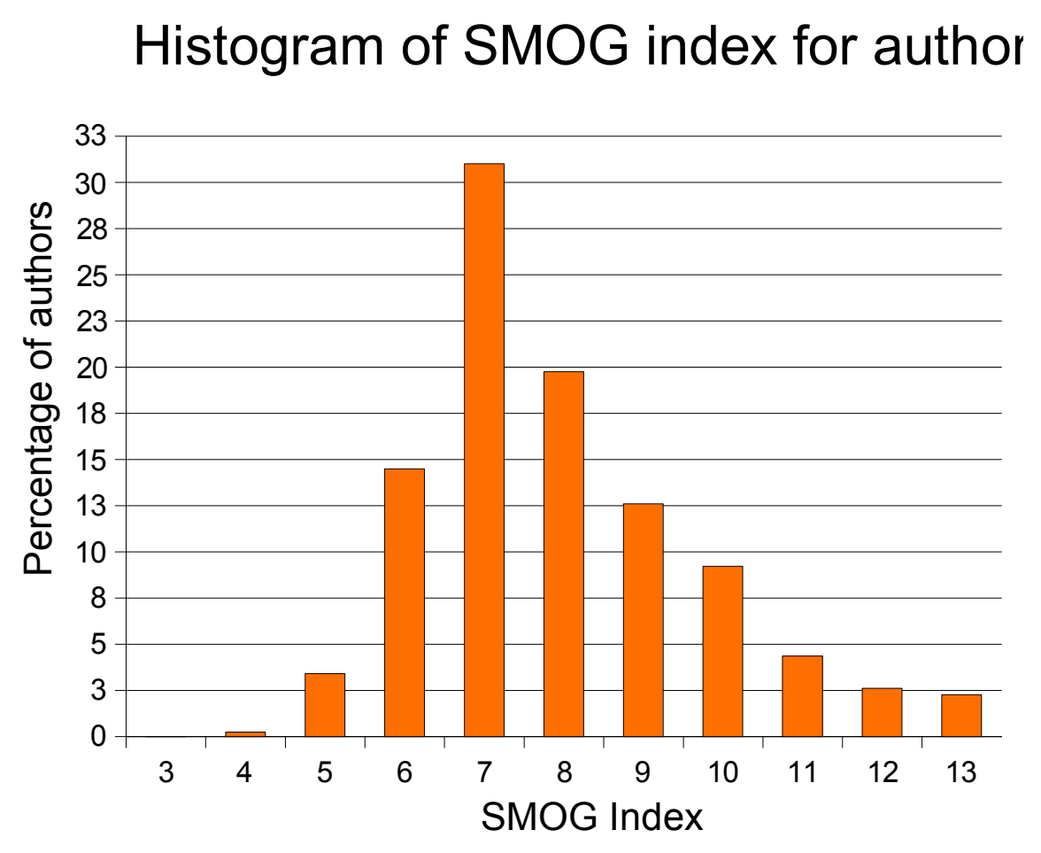
The data was collected from USENET newsgroup servers as follows. The list of newsgroups provided by the server was obtained and groups with "binaries" in their name were manually removed. All available articles from the comp.* and sci.* hierarchies were saved in a relational database. Articles were them drawn from groups in the alt.* hierarchy until approximately 3.1 million Usenet articles were retrieved. The posting dates of these articles span approximately 139 days.

The headers of each article were parsed in order to obtain the author's provided email address, the article's subject, the date of posting, and references. This information was then used to compute measures relating this article to others, such as the article's depth within the article tree, the article's number of siblings, and its number of immediate children. Various readability measures were also computed for the contents of the article itself, including the SMOG readability Index.

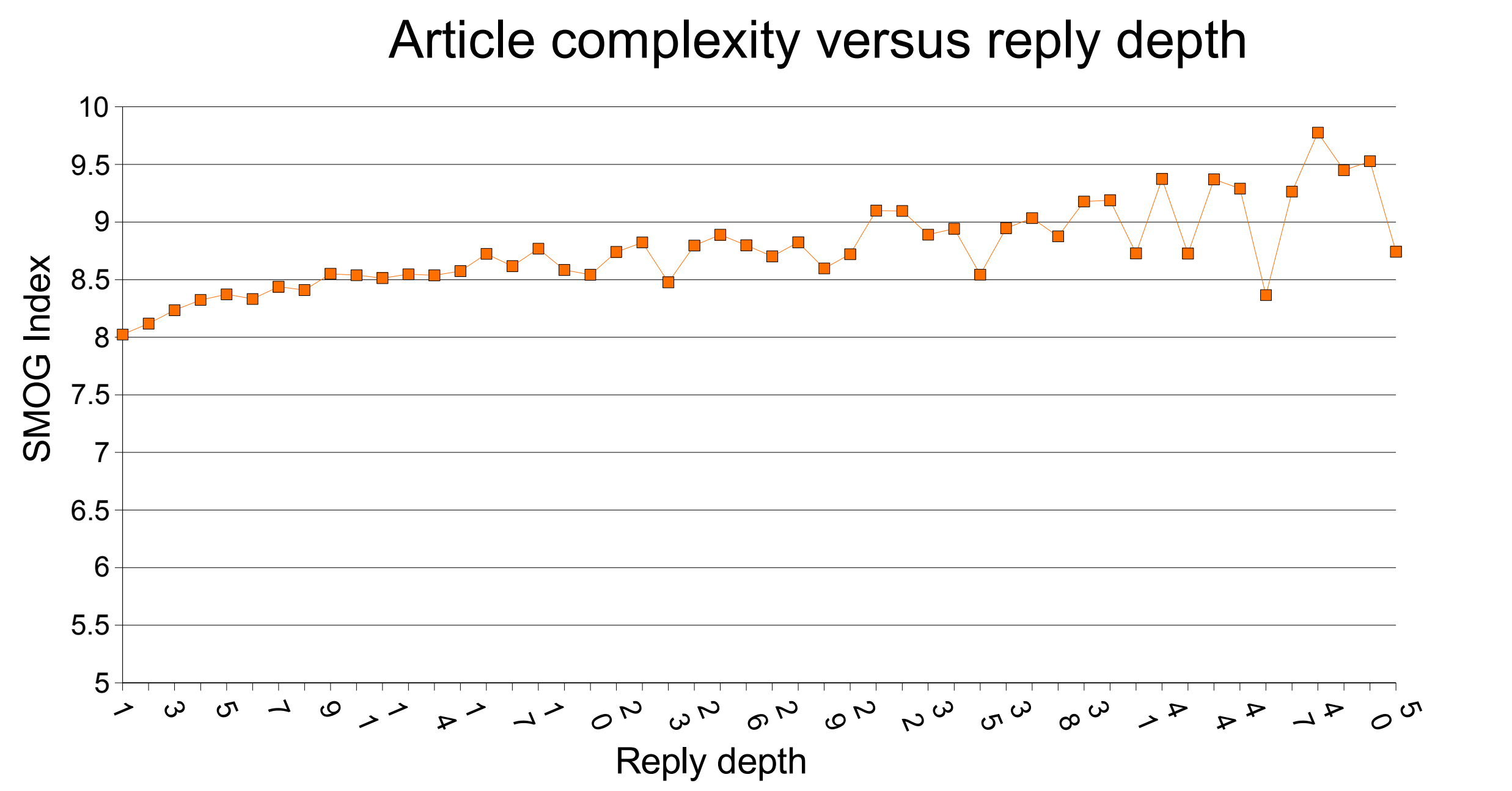
Authorship complexity

The SMOG readability index returns a grade level based on the complexity of the underlying text. Overall we found that the average SMOG index for the average author was 8.4. However, the average SMOG index for all articles was slightly higher at 9; which would suggest that the slightly more educated authors publish more often.

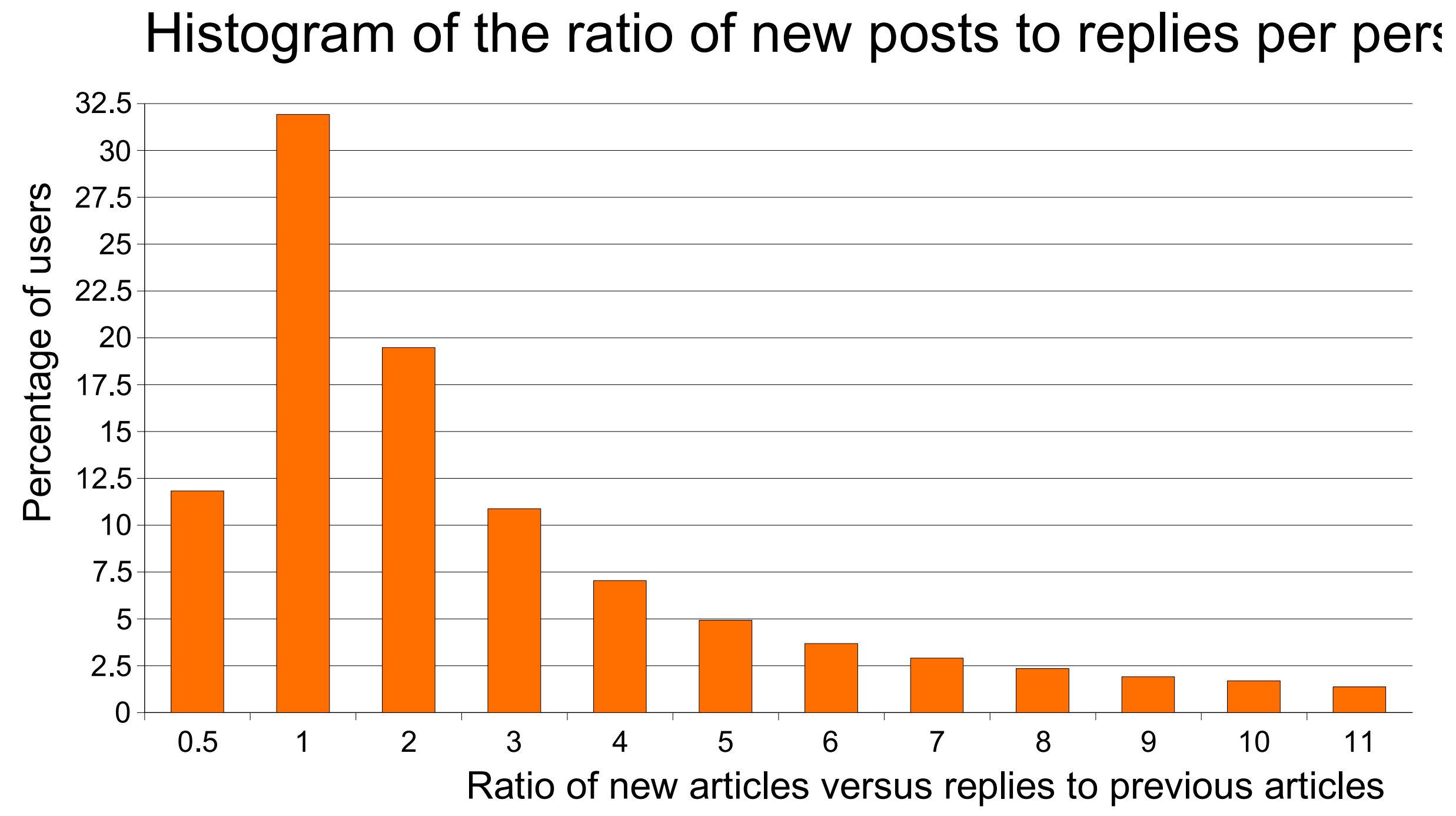




Complexity and Replies



A question that we wished to have answered was about the complexity and the quality of online discussion groups. Specifically, do discussions degenerate into squabbling over minor points or do they bring out serious debates involving complex ideas?
The above graph seems to indicate that the discussions are on average healthy as the average SMOG index increases as the length of the discussion progresses.



Similarly, we plotted a histogram of the ratio of new topics posted by authors to the number of replies to old topics. Overall, the behavior of the users seems to be balanced in between suggesting new topics and contributing to existing discussions.

The mythical Troll

A phenomenon that is common to Internet lore is the concept of the 'troll' that will post inflammatory and provocative comments to a discussion. We did find ample evidence of this kind of behavior and were able to build models to detect these behaviors by measuring the rate of new replies within a 24 hour period. In our tests, we used the arbitrary threshold of 100 replies/day to locate this kind of behavior. This is useful for an algorithmic perspective since it serves to identifies discussions with little or no new information for future parsing.

Topic	Explosive Threads
alt.atheism	6
alt.impeach.bush	2
talk.politics.guns	2
alt.rush-limbaugh	2
alt.gossip.celebrities	2

Topic Complexity

Topic	Smog Measure
sci.physics	8.81
alt.society.liberalism	8.59
us.military.army	8.37
alt.religion.christian	8.25
alt.politics.bush	8.13
alt.atheism	8.1
alt.politics.republicans	7.94
alt.coffee	7.87
talk.politics.guns	7.77
alt.battlestar-galactica	7.54
alt.astronomy	7.52
alt.fan.harry-potter	7.48
alt.home.repair	7.38
alt.fashion	7.16
alt.gossip.royalty	6.97
alt.fan.howard-stern	6.91
rec.sport.pro-wrestling	6.63
alt.arts.poetry.comments	6.51
us.arts.poetry	6.35
alt.poetry	6.28
alt.fiftyplus	6.25
alt.tasteless.jokes	5.64
alt.funnytown	5.61
rec.humor	5.49
alt.humor	5.42
mn.humor	5.38
alt.humor.puns	5.18

We computed the average SMOG index for a sample of the collected newsgroups and ranked them according to the index value. The results are for the most part intuitive, except however for the entertainment discussion groups "alt.battlestar-galactica" and "alt.fan.harry-potter". Our hypothesis is that these topics make use of a more complex vocabulary because of the nature of the fantasy world they discuss.

Social distance

Topic	Social distance	Connectivity
alt.religion.christian	3.97	0.87
alt.politics.bush	3.59	0.96
rec.humor	3.57	0.77
talk.politics.guns	3.49	0.81
alt.society.liberalism	3.44	0.95
sci.physics	3.41	0.96
alt.home.repair	3.38	0.97
alt.politics.republicans	3.38	0.93
rec.sport.pro-wrestling	3.36	0.98
alt.humor	3.35	0.56
us.military.army	3.33	0.90
alt.fan.howard-stern	3.31	0.93
alt.astronomy	3.26	0.94
alt.atheism	3.19	0.98
alt.coffee	3.15	0.98
alt.fashion	3.06	0.87
alt.fan.harry-potter	3.05	0.97
alt.arts.poetry.comments	3.00	0.97
alt.tasteless.jokes	2.97	1.00
alt.battlestar-galactica	2.93	1.00
alt.hvac	2.88	0.99
alt.gossip.royalty	2.80	0.85
alt.humor.puns	2.54	1.00
alt.fiftyplus	2.34	1.00
alt.funnytown	2.06	0.95
mn.humor	1.96	1.00

For a number of sampled newsgroup we computed the social distance (e.g.:Whether a discussion thread occurred that linked the individuals) and connectivity (The probability that two individuals are somehow connected). The results we obtained are similar to those that we expected, yet we were surprised at how integrated discussion forums were even in high volume discussion groups (alt.politics.bush has over 33,000 articles)..

Similarly, small, specialized communities such as alt.hvac (Heating, Ventilation and Air Conditioning) behaved as very highly integrated, almost fully connected communities that are very efficient at transporting information. The overall low social distance in all cases is especially surprising considering the volume of information and the size of the social groups.

Conclusion

Through the use of structural and readability metrics, we are able to locate and identify a number of social events and characteristics that may be helpful in understanding and analyzing texts. In later work we will make use of these metrics to create parameters for parsers in Information Retrieval engines.

Acknowledgments

The authors would like to thank Ian Mackinnon and Jonathan Fishbein.