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 structure 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 "binary" in their name were manually removed. All available articles from the comp.* and sci.* hierarchies were saved in a relational database. Articles were then 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

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

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 found 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 identify discussions with little or no new information for future parsing.

The above graph seems to indicate that the discussions are on average debates involving complex ideas?

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.