Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Hierarchy Prediction in Online Communities

Denys Katerenchuk[⋆] and **Andrew Rosenberg**[⋆†]
dkaterenchuk@gradcenter.cuny.edu, andrew@cs.qc.cuny.edu
[⋆]CUNY Graduate Center, New York, USA
365 Fifth Avenue, Room 4319, New York, NY 10016
[†]CUNY Queens College, New York, USA
65-30 Kissena Boulevard, Room A-202, Flushing, NY 11367

Abstract

With the development of the Internet, a big part of social interactions have moved online, and people have unconsciously brought their daily communicational habits to the web. Understanding these communications is important because it will lead to a better understanding of online communities, and can improve areas such as e-commerce, advertisement, topic modeling, security, and others. We propose to develop a natural language based ranking algorithm to predict user influence levels in online communication groups.

1 Introduction

Communication is a part of our existence. Through interactions we build social hierarchies. Scientists have shown that people change their communicational behavior based on their collocutor (Yaman, Hakkani-Tür, and Tür 2010). These changes can reveal a lot of information about a conversation's participants. In recent years, a big part of interactions has moved to the Internet. Inevitably, this leads to a creation of hierarchical relationships similar to those formed during face-to-face interactions. Understanding user relationships and social structures are essential to many areas such as e-commerce, advertisement, security, and others.

We propose an algorithm to improve current hierarchy prediction from text communications, a well known problem in natural language processing (NLP). While early studies were concentrated on social network analysis (SNA) (Diesner and Carley 2005), in recent years multiple studies have been done on text communication analysis (Gilbert 2012). Still, the hierarchy prediction task remains a challenge. In this paper, we propose to bring together different techniques from NLP to improve the current state-of-the-art algorithms for hierarchy prediction from text.

2 Related Work

With the release of Enron email corpus (Klimt and Yang 2004), many scientists have tried to predict corporation structure from email communications. Initial effort was focused on SNA algorithms (Diesner and Carley 2005). On the other hand, more work has been done in recent years on text analysis. The researchers studied word usage (Bramsen et al. 2011) and showed that simple word frequencies can be indicative of power relationships. They used simple n-grams, part-of-speech n-grams and mixed n-grams models in their experiments. Eric Gilbert also used n-grams in his research (Gilbert 2012), but conducted more detailed text analysis. All of these studies considered dyadic communications. In our research, we consider communicational threads where participants reply to multiple users, and we rank the users according to their hierarchy. Ranking user groups makes the task much harder, but provides broader background for understanding user communities.

3 Data

A big part of research on hierarchy prediction is done on Enron corpus and wiki talk pages ("talk" tab). We, however, want to perform analysis on data that is closer to real world communications. For this reason, we collect data from Reddit[1]. Reddit is a website where users create posts on different topics or share resources such as pictures, videos, or links to other resources. Users can participate in discussions through creating threads of comments. Each comment can earn comment karma, which is Reddit's form of approval. The comments and posts can be on any subject and not restricted to follow any rules such as business style emails in Enron data or Wikipedia improvement discussion. However, to facilitate discussions, we collect data from politics subreddit.

We define user hierarchy in terms of k-index. The definition of k-index is influenced by familiar author-metric index, h-index, used in science. K-index means that a user has k posts with at least k karma scores. We assign k-index to each user inside a single thread. This means that a user can have high hierarchy level in one thread, but low in another. In other words, we consider only local user hierarchical levels.

4 Methods

In this paper, we propose to develop an algorithm that brings together a state-of-the-art NLP research from multiple subareas to enhance hierarchy prediction in online communications. Our proposed algorithm relies on data from different NLP domains.

[1] www.reddit.com

4.1 N-grams

To start, we base our algorithm on well known and established word and part-of-speech n-gram model. Multiple studies have shown effectiveness of this approach (Bramsen et al. 2011; Gilbert 2012). In addition, we consider function words discovered by James Pennebaker shown to carry information about an author (Pennebaker, Mehl, and Niederhoffer 2003).

4.2 Word Length

Users of different hierarchy vary in word choice. In particular, users with higher hierarchical status use more complicated words (Dino, Reysen, and Branscombe 2008). We believe that taking word length into account will help distinguishing hierarchies.

4.3 Politeness

Politeness in conversations is one way to see the dominance. From a study of StackExchange, users with lower status were found to be more polite (Danescu-Niculescu-Mizil et al. 2013). In fact, as soon as they reach high status, they stop being polite and become offensive. Sentiment analysis can bring more information to understand text communications.

4.4 Hedging

A linguistic hedge is a device that is used to mitigate the meaning of a statement, request or question. Hedges are somewhat controversial. Studies have shown that hedges are used to make participants sound polite and can be indicative of lower status participants (Lakoff 1975). While the other studies show that hedges used by authors to distant themselves from direct orders (Skelton 1988). Hedges are important and can further improve our algorithm.

4.5 Entrainment

Entrainment is a way collocutors mimic each other's style of communication (Watanabe, Okubo, and Kuroda 1996). We believe that users with high status influence conversations. They can bring different writing style, new words or change a topic.

We want to use word embedding method for this task (Mikolov et al. 2013). Word embedding is a way to represent words in n-dimensional mathematical space. The model is trained using deep neural networks. After training, each word in the model has n-dimensional representation. Interestingly, the words that are close in meaning are close to each other in n-dimensional space. In this way, we can measure similarities among comments.

4.6 Classification

Since the task is to predict hierarchies in a thread, the predictions are relevant to a given subset of users. For this reason, we compare and analyze different machine learning ranking techniques such as support vector regression machines (Smola and Vapnik 1997), IntervalRank (Moon et al. 2010), and other ranking algorithms.

5 Conclusion

In this paper, we propose a unique algorithm that combines multiple theories from NLP and linguistics to improve hierarchy prediction. Many of proposed theories have shown to be helpful, but never been used together. Combining these theories into a single model will improve the best hierarchy prediction task algorithms.

References

Bramsen, P.; Escobar-Molano, M.; Patel, A.; and Alonso, R. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 773–782. Stroudsburg, PA, USA: Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Diesner, J., and Carley, K. M. 2005. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*. Citeseer.

Dino, A.; Reysen, S.; and Branscombe, N. R. 2008. Online interactions between group members who differ in status. *Journal of Language and Social Psychology*.

Gilbert, E. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1037–1046. ACM.

Klimt, B., and Yang, Y. 2004. Introducing the enron corpus. In *CEAS*.

Lakoff, G. 1975. *Hedges: a study in meaning criteria and the logic of fuzzy concepts*. Springer.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moon, T.; Smola, A.; Chang, Y.; and Zheng, Z. 2010. Intervalrank: isotonic regression with listwise and pairwise constraints. In *Proceedings of the third ACM international conference on Web search and data mining*, 151–160.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Skelton, J. 1988. Comments in academic articles.

Smola, A., and Vapnik, V. 1997. Support vector regression machines. *Advances in neural information processing systems* 9:155–161.

Watanabe, T.; Okubo, M.; and Kuroda, T. 1996. Analysis of entrainment in face-to-face interaction using heart rate variability. In *Robot and Human Communication, 1996., 5th IEEE International Workshop on*, 141–145. IEEE.

Yaman, S.; Hakkani-Tür, D.; and Tür, G. 2010. Social role discovery from spoken language using dynamic bayesian networks. In *INTERSPEECH*, 2870–2873.