This exemplar, prepared by RUCS editors, is an extended abstract of a 2013 research paper published in the Stanford Undergraduate Research Journal. The full paper can be found here.

Predicting Edit Locations on Wikipedia using Revision History

Caitlin Colgrove, Julie Tibshirani, Remington Wong

Introduction

In the machine learning community, there has been increasing interest in automated task design. In a collaborative problem-solving setting, how can we best break up and assign tasks to optimize output?

Back in 2007, Cosley et al. created SuggestBot to help Wikipedia contributors find articles of interest in Wikipedia's massive repository [1]. Using data from a contributor's revision history, SuggestBot found related articles contributors may want to work on next, which quadrupled edit rates by users. In an effort to give more specific, polished suggestions and increase contribution rates even further, we considered an extension of this problem: once a Wikipedia user has chosen an article to view, which *section* will they most likely be drawn to?

Following the approach used in the creation of SuggestBot, we hypothesized that users' revision histories, together with the most predictive characteristics of a section, help determine which section will be next edited. Unlike SuggestBot, machine learning was used in making the predictions.

Approach

The core of the approach was to use a single editor's revision history to train a Naive Bayes model, a probabilistic classifier that, given some input, creates a probability distribution over a set of classes to which that input could possibly belong. It is naive because it makes assumptions of statistical independence – that is, that no event has any effect on the probability of any other.

To make a prediction on a test article, the classifier calculated the likelihood of each section being edited given lexical features drawn from a user's edit history and simple features of the section itself. Then, the sections were ranked by likelihood and these ranks were outputted. Initial results showed that the classifier overwhelmingly favoured short paragraphs, ostensibly because longer sections were saturated with uninformative words; we subsequently corrected the model for this.

Since revisions are typically very short, they did not provide ample training data. The edit histories that were downloaded, however, contained both the lines that were edited and a few lines surrounding the change, which we then included to increase the quantity of training data. A second modification was the use of the entire text of the edited article, which allowed the classifier to further characterize the edited section by identifying words unique to it. Words that appeared elsewhere in the article could therefore be ignored.

Analysis

Defining success as the inclusion of the correct section in the top three ranked by the classifier, we contend that the algorithms performed very well. In most cases, the baseline, which made predictions based solely on section length, predicted the correct section 80% of the time. For 40% of users, this baseline provided the best results. Using Naive Bayes sometimes made a marginal improvement on this baseline, and, 20% of the time, none at all. Experimenting with the inclusion of an edit's surrounding text also yielded little improvement.

Despite a satisfactory overall performance, different models seemed to perform better on different users. Naive Bayes worked well on users who consistently edited the same sections

of each article they read. For users with localized expertise, however, who changed different sections of closely-related articles, a high degree of semantic similarity between their revision histories and the articles they tended to work on made searching for a trend in their edit history ineffective. Making accurate distinctions between previous revisions whose contents have a significant overlap with the entire article proved difficult.

Conclusion

We achieved a surprisingly good performance with the baseline alone, which considered only the length of each section. Moreover, it would seem that the characteristics of a section have far more predictive power than a user's edit history. This is ostensibly why including the text surrounding edits did not make predictions noticeably more accurate. The findings also lead to the somewhat surprising conclusion that most Wikipedia users are copy editors as opposed to content editors.

By building off the work of Cosley et al.'s SuggestBot, we have identified a method for making more fine-tuned recommendations to Wikipedia contributors. In doing so, this paper not only marks an improvement with an existing case study, but it also demonstrates a more nuanced application of automated task design through machine learning, setting a precedent for similar such applications in the future.

References

1. Cosley D, Frankowski D, Terveen L, et al, "SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia" in Proc. 12th Int. Conf. in Intelligent User Interfaces., Honolulu, HI, 2007.
2. Klein D, Manning C, "Accurate Unlexicalized Parsing" in Proc. 41st Meeting Association for Computational Linguistics., 2003.
3. Pennacchiotti M, Popescu A, "Republicans and Starbucks Afficionados: User Classification in Twitter" in Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, New York, NY, 2011.