Modelling Discussion Topics to Improve News Article Tagging

Chris Emmery
CLiPS, University of Antwerp
chris.emmery@uantwerpen.be

Menno van Zaanen
TiCC, Tilburg University
mvzaanen@uvt.nl

Online news articles are often labelled by their writers with a set of tags to topically frame their content. For many news websites, the tags allow for easier retrieval of articles by news readers after the articles have vanished from the front pages. In addition, tags cluster articles by topic, granting the ability to quickly search through multiple articles on the same topic.

We show that human-provided tags have two major shortcomings. Firstly, they often result in several uninformative tags which are seldom reused. Secondly, the social context of the article is not always fully described by the tags. Readers may have associations or ideas in relation to an article that were not predicted by its writer and, hence, corresponding tags are missing. The first problem can be resolved by frequency filtering of the tags. We argue here that the second problem can be resolved by detecting the latent topics in online discussions that directly relate to the article (for instance, discussions in the comment section below news articles) and link them to the tags of the article that were provided by the writer. The intuition behind this idea is that information taken from the context of previously written articles can be leveraged to improve the quality of the tags. Additionally, information from the previously identified latent topics can be reused to assign tags to new articles.

In this research, we employ a supervised topic model to a collection of news articles and their associated comments to detect latent topics. By utilizing the tags that are already assigned to the articles, we are able to train a Labeled Latent Dirichlet Allocation (L-LDA) model in a supervised setting. By doing so, we can evaluate tags proposed by the fitted model for unseen articles. This is done in a ranked setting using Mean Average Precision (MAP) against the original human-provided tags.

We will demonstrate the influence of the social context of a news article, taken from the related discussion, on the performance of the overall model. Additionally, we can investigate the quality of the model by comparing its inferred individual tags to those of the writer. Finally, this approach allows for a close collaboration between the system and the writer of news articles to improve tag assignment, with the aim of improved searches for related articles.