Tag Recommendations for StackOverflow Posts

CSE 6240 Project Proposal
Ke Wang 903058889
Haochen Zhao 903070441

Abstract

Our team plan to work on tag recommendations for posts on StackOverflow website.

Introduction and User Need

As we have seen on Twitter or Facebook, tags or hashtags are short labels, provided as metadata. By introducing the idea of tags, we can build indexes on them, make categories, and enable personalized bookmarks. One of the most popular software information sites, StackOverflow, supports tags as well. However, StackOverflow doesn't provide tag recommendations yet, and at the same time, a user must provide at least one tag when writing a new post. Even experienced users may sometimes have no idea what tags he can provide. So our goal is to recommend tags to users. Basically, our recommendation is based on what a user has written in his new post. We plan to analyze the contents of the posts, combine information from history posts with tags, and recommend most likely tags to users.

Solutions

We have found a couple of ideas from some papers, which we can probably use in our project. The solutions can be roughly divided into two categories. One is based on latent semantics, and the other is based on term frequency. As solutions based on latent semantics are expected to be more powerful than those based on term frequency, we are mainly focusing on LDA-based solutions. [4] gives a LDA-based solution that after we calculate all parameters, we can get the probability of each possible term by $P(c|d,\alpha,\beta) = \sum_{t=1}^{T} P(c|t,\beta)P(t|d,\alpha)$. Then we return the most likely term as recommended tags. [1] proposes an idea that because a document is a distribution over topics, the similarity of two texts can be computed in terms of similarity of distributions. We can make use of this idea. After we compute the similarity between the current post a user has just written and history posts, we can regard history posts with best similarity as neighbors of the current post, and we can recommend the tags in these history posts. There are many ways to calculate distribution similarity. [4] provides a few like *Kullback-Leibler divergence* and *Information Radius.* [2] provides more.

Term frequency based methods are often more intuitive, but less accurate. For example, TF-IDF can be used to solve our problem. We compute TF-IDF value for each of

possible keywords in the post, sort them by TF-IDF value, and return top keywords as recommended tags. [3] combines three solutions. One interesting solution among the three is to observe from history posts tag-term co-occurrence. Then derive tags that have highest probability to appear.

Our group plan to mainly focus on LDA-based methods, and implement some term frequency based methods as well to compare these methods, and maybe we can combine these methods, like [3] does to get a better result.

Expected application

Our expected application of our tag recommendation is that on websites such as StackOverflow, when a user writes a new post, we can automatically recommend some tags to the user based on what the user has written and history posts. Then a user can choose from these tags efficiently.

Datasets

Our goal is to recommend tags for StackOverflow posts, so we will mainly use StackOverflow datasets. StackOverflow provides dump files, which contain most of what we need such as existing tags and history posts. Besides, we may also apply our implementation to AskUbuntu posts.

References

[1] Rus, V., Niraula, N., & Banjade, R. (2013). Similarity measures based on latent dirichlet allocation. In *Computational Linguistics and Intelligent Text Processing*(pp. 459-470). Springer Berlin Heidelberg.

[2] Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.

[3] Xia, X., Lo, D., Wang, X., & Zhou, B. (2013, May). Tag recommendation in software information sites. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (pp. 287-296). IEEE Press.

[4] Dredze, M., Wallach, H. M., Puller, D., & Pereira, F. (2008, January). Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces* (pp. 199-206). ACM.

[5] Zhu, T., & Li, K. (2012). The Similarity Measure Based on LDA for Automatic Summarization. *Procedia Engineering*, *29*, 2944-2949