Improving GitHub Tag Recommender Systems Using Tag Hierarchies

1: Background and Goal

- GitHub repositories can be assigned tags or topics
- These support search queries, which is useful
- Tag recommender have already been developed, without hierarchy[1]
- See if recommending tags using a hierarchy is better than not using a hierarchy.

2: Approach

- First, we collect Hierarchical Multilabel Classifiers (HMCs)
- Next, we create a hierarchical structure for the tags
- Then, we train the HMCs with the hierarchies
- Finally, we compare performance between a baseline and the best performing HMC, using AUPRC

Contact details

Arend van der Rande
a.c.vanderrande@student.tudelft.nl

3: HMCs

- Are a type of classifier that can assign multiple labels to an item
- They use hierarchical information to improve recommendations

3.1: AWX

Uses an output layer of a neural network with a special loss function

3.2: C-HMCNN(h)

Also uses an output layer, but with a hierarchical loss and constraint function

3.3: HMC-LMLP

Is a stack of neural network, each predicting a layer of the hierarchy.

3.4: HMCN-F

Is an extension of HMC-LMLP, with the input features also giving input to each layer and a global loss function.

4: Hierarchies

- For creating the hierarchies, we use clustering algorithms: bisecting K-means and agglomerative clustering.
- These algorithms need a distance metric between tags, for which we use the SED-KGraph[2] and a co-occurrence matrix
- This results in four hierarchies: SEDK-BK, SEDK-AC, COM-BK and COM-AC.
- The spread of high-level cluster sizes can be seen in Figure 1

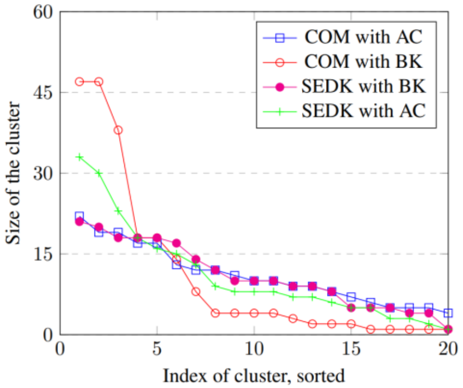

Figure 1: Comparison of cluster sizes

5: Results

The classifiers are compared against the baseline, LR in AUPRC scores. Trained on 10000 repositories and 220 tags.

	SEDK-BK	SEDK-AC	COM-BK	COM-AC	LR
AWX	0,539	0,542	0,546	0,542	0,556
C-HMCNN(h)	0,373	0,372	0,355	0,357	-
HMC-LMLP	0,121	0,107	0,091	0,128	-
HMCN-F	0,564	**0,570**	0,568	0,556	-

Table 1: AUPRC scores from the HMCs (left) combined with the hierarchies(top)

6: Conclusion

- HMCs can outperform the baseline
- However, currently this is marginal
- Potentially, a different construction for hierarchies HMCN-F to outperform LR by a significant margin

References

[1] M. Izadi, A. Heydarnoori, and G. Gousios, “Topic recommendation for software repositories using multilabel classification algorithms,” Empirical Software Engineering, vol. 26, no. 5, p. 93, Sep. 2021

[2] M. Izadi, and A. Heydarnoori, “Semantically-enhanced Topic Recommendation Systems for Software Projects,” Tech. Rep., 2022.