Category suggestion for e-commerce queries
L. Mooiman

Abstract

This thesis proposes a category suggestion model for an e-commerce search engine to get an insight on the intent of the visitors. The dataset is provided by the Dutch e-commerce website Bol.com and consists of visitors queries and their subsequent clicks on items. The first part of the model is clustering the dataset to decrease the complexity and find previously unseen similarities between queries. Each query is weighted to all the other queries with a combination of the features lexical matching, word2vec, TF-IDF, co-clicks and a category tree match. The maximum weighted query pair clusters together and the model keeps clustering until it reaches a stop condition. The second part is creating a category distribution for every cluster from the clicks on the items associated with specific categories.

The model is then tested by matching a query from the test set to all the clusters with a combination of the two features lexical matching and word2vec. The query inherits the category distribution of the cluster with the highest weight, and it is accurately assigned when the category of the query matches the top category of the inherited distribution. The whole test set is evaluated and results in the accuracy of the model.

Various models are created with different combinations of the features and stop conditions. Each is evaluated with the same test set on their accuracy when the test features are varied as well, and the attributes of the clusters of each model are discussed. As are the general observations of all the models.

The results vary depending on the importance of accuracy or speed. The model with the highest accuracy uses only the feature word2vec, and the test features is a combination of 70% lexical matching and 30% word2vec, which results in 66% accuracy. However, word2vec is slow, and the fastest model is the same model but only with the test feature lexical matching. The top 1000 queries are thus given categories suggestions, and only one cannot be assigned a category.

Observations are made of a few large clusters in the model and the large quantity of one-query clusters. The aspects of the features are also discussed, such as the sparsity of the co-clicks feature. Future work includes adding more features to minimize the number of one-query clusters and optimization of feature selection.

In conclusion, the category suggestion model returns proficient category distributions for an unseen query and an insight into the visitors' intent.