Learning to Classify Short and Sparse Text Web with Hidden Topics from Large-scale Data Collections

摘要

本文提出了一个用于构建分类器的通用框架，通过充分利用从大规模数据集合中发现的隐藏主题来处理稀疏的短文本。该研究工作的主要动机是，由于数据稀疏性，许多针对短文本的分类任务都不能实现高准确率。因此我们提出了一种想法，即通过获取外部知识来让增强数据的相关性，同时扩大分类器的覆盖率，从而更好地处理未来的数据。该框架的底层思想是，对于每一个分类任务，我们收集大规模的外部数据集合，然后在两个数据集——（小的）标记训练数据集和从上述的大规模数据集合中发现的隐藏主题数据集——上构建一个分类器。该框架具有良好的通用性，可以被应用到不同的数据域。

6.1 Choosing Machine Learning Method

列举了一些分类方法，比如 K-NN, Decision Tree, Naive Bayes, MaxEnt, SVM；并给出了本文选择 MaxEnt 的原因。