博士论文

基于句类向量空间模型的自动文本分类研究

张运良1,2，张 全2

(1. 中国科学院研究生院，北京 100039；2. 中国科学院声学研究所，北京 100080)

收稿日期 修回日期 网络版发布日期 2007-11-16 接受日期

摘要 向量空间模型是自动文本分类中成熟的文本表示模型，通常以词语或短语作为特征项，但这些特征项通常只能提供较少的局部语义信息。为实现基于内容的文本分类，该文用HNC理论中的句类作为特征项，通过混合句类分解等技术对句类向量空间降维，使用tfc算法对特征项进行权重计算，用KNN算法进行分类。该分类器的平均准确率和召回率都是可接受的，对类别的抽象程度无要求，即抽象度较高和较低的类别可以同时分类。通过使用更好的机器学习算法和其他的HNC语言理解技术，性能可以进一步提高。

关键词 文本分类 句类 向量空间模型 HNC理论

分类号 TP391

DOI:

通讯作者:

作者个人主页: 张运良1;2;张 全2