

评“半监督语义动态文本聚类算法”

说起聚类，第一反应是“物以类聚，人以群分”，同类的东西常聚在一起，志同道合的人相聚成群。在聚类问题场景描述中，我们通常认为参与聚类的这些数据是事先存在的、定量的。然而，在现实生活中，数据并不确定，而是动态的、增长的、变化的。在这样的数据环境下，聚类方法也必须适应数据的变化。反映在聚类中，各类别所包含的数据样本会不断增长和改变，甚至类别的含义也会随着时间的变化而发生改变。为此，我们需要动态的文本聚类算法，来捕捉动态环境下聚类结果的变化和发展。

我们再把眼光放到语义学习方法的发展上。远到 word2vec 模型的提出，近到 BERT 模型的问世，都引领了新一批语义学习的研究工作。语义学习方法大大提升了实体抽取、情感分析、语法分析等众多自然语言理解和文本挖掘任务的效果。这些方法为动态文本聚类算法的研究注入了活力，为动态文本聚类中类别语义的表示带来了新的思路。但是，如何更好地将语义学习方法融合于动态文本聚类中，仍然面临许多问题，比如，如何解决聚类工作中一直面临的难题——准确估算聚类个数？如何让类别的含义紧跟数据的变化？

该文作为一篇将语义嵌入与动态文本聚类相结合的文章，提出的半监督聚类方法更好地捕获了聚类结果，具有一定的创新性和良好的应用前景。相信该文会对研究动态文本聚类的学者带来思考和启发。

评“基于上下文语义的新闻人名纠错方法”

随着移动互联网的快速发展，人们对新闻时效性的要求日益增长，但新闻的真实性和准确性仍然不可忽视。如时政新闻中的人名书写错误，就会在新闻的传播过程中带来诸多不良影响。因此，如何利用文本分析技术自动发现和及时纠正新闻文本中的人名错误，是新闻行业中亟待解决的一个新问题。该文提出的基于上下文语义的新闻人名纠错方法正是针对该问题提出的一种有效的解决方法。

传统的纠错算法常用到编辑距离，通过计算编辑距离的差异来纠错。这种方法只考虑了人名书写相似性这一单方面的信息，但在文本中除了书写相似以外，还有很多其他值得我们考虑的信息，如：人名上下文中出现的人物重要属性、人物在新闻中的常见描述背景等。如果读者也在思考如何有效地利用更多的信息进行人名纠错，那你将在这篇文章中得到启发。

该文提出的基于上下文语义的人名纠错方法，针对疑似人名错误，可融合人名的语义环境特征、关键属性特征，人名的拼写和字形特征三方面特征，自动计算判别错误人名。此外，该方法亦可认为是命名实体识别的改进，可通过融合上下文语义识别包含疑似错误字符的人名。