从问答讨论中发现语义相关的技术术语和网络资源

贾俊芳[1]，Valeriia TUMANIAN[2]，李国强[2]
[1]山西大同大学计算机与网络工程学院，中国大同市，037009
[2]上海交通大学软件学院，中国上海市，200240

摘要：目前网络上拥有大量可用于软件工程实践的技术和网络资源，并且这个数量还在持续增长。发现语义相似或相关的技术术语和网络资源，可以设计吸引人的服务，以促进信息检索和信息发现的机会。本文从问答（Q&A）讨论的社区中提取技术术语和网络资源，并提出一种基于神经网络语言模型的技术术语和网络资源在联合低维向量空间中的语义表示方法。方法仅基于讨论线程中技术术语（或网络资源）的周围技术术语和web资源，将技术术语和网络资源映射到语义向量空间，而不需挖掘讨论的文本内容。将方法应用于2018年3月的堆栈溢出数据转储。对聚类、搜索和语义推理任务的定量和定性分析表明，所学习的技术术语和网络资源向量表示可以捕获技术术语和网络资源的语义相关性，通过简单的*K*近邻搜索和在嵌入空间中对学习的向量表示作简单的代数运算，可以支持各种搜索和语义推理任务。

关键词：技术术语；网络资源；词语嵌入；问答网站；聚类任务；推荐任务
https://doi.org/10.1631/FITEE.2000186