doi:10.1631/FITEE.1500187

题目：基于共同共现群体相似度的社会化标签聚类方法

目的：社会化标注系统产生了大量具有歧义和非受控的标签，降低了用户体验也限制了资源检索效率。标签聚类能够将具有相似语义的标签聚集在一起，从而缓解上述问题。现有的社会化标签聚类方法基本上从“资源-标签”的二元关系测量标签相似度，并使用 *K*-means 和层次聚类等算法实现标签的聚类，容易引起高维、稀疏和标签语义丢失等问题。本文提出一种基于共同共现群体的标签相似度测量方法，利用谱聚类算法实现标签聚类。

创新点：对社会化标注系统中的三元标注关系进行分析，总结出三元关系中最能保持语义关系的标签共现形式。在分析标签个体共现相似度的基础上，利用群体思想，提出标签的共同共现群体相似度，从全局角度精准地刻画标签的语义相似性，并提出一种基于共同共现群体相似度的社会化标签谱聚类方法。

方法：利用共同共现群体相似度来计算两两标签的相似度，建立相似度矩阵（公式（4））。使用谱聚类算法实验标签的聚类，首先使用拉普拉斯（Laplacian）变换对相似度矩阵进行规范化，建立标签的规范化拉普拉斯（Normalized Laplacian）矩阵，然后计算该矩阵的前 k 个特征值及其对应的特征向量，并将这 k 个特征向量组成新的特征空间，在此空间上用 *K*-means 算法将标签聚成 k 个类簇（算法 1）。

结论：利用内部评价指标 SC 和 Dunn 对本文提出的标签聚类方法和其它传统的标签聚类方法进行实验对比。得出基于共同共现群体相似度的标签谱聚类方法在 SC 和 Dunn 这两个指标上的值均优于其它传统标签聚类方法；基于共同共现群体相似度的标签谱聚类方法能够获取较好的聚类结果。

关键词：社会化标注系统；标签共现；谱聚类；群体相似度