# Document Clustering based on Topic Maps

**摘要**

文本聚类的两个关键问题：(1) 在文档表示上，需要捕获文本中的语义信息——这对文档的降维表示会有帮助；(2) 基于文本的语义表示定义相似性度量，使得语义关联度较高的文档之间具有较高的相似度。文档的特征空间会对文档聚类造成很大挑战。一个文档可能包含多个主题，以及大量类别独立的通用单词 (即 stopwords)、类别特定的核心单词 (即该类别文档的 keywords)。这使得传统的基于 Document Vector model 或 Suffix Tree model 的凝聚式 (agglomerative) 聚类算法无法获得较高的聚类质量。本文提出了一种新的文档聚类方法，该方法基于文档的主题映射 (Topic Map) 表示，将文档转换为一种密集形式 (compact form)。文档的相似性则是基于从主题映射的数据和结构中所推断出来的信息进行度量的。

## Document Clustering Based on Topic Maps

文档聚类包括 3 个基本步骤，我们的聚类算法也遵循这些步骤。我们首先将每个文档转换为密集形式 (compact form)，该密集形式仅表示文档中出现的主题以及主题之间的 occurrence 和关联。topic maps 信息使用 Wandora[14] 生成。使用 Wandora 的导出功能可以将 topic maps 导出为 XTM 格式。完成对 topic maps 的收集后，可以生成一个相似性度量，用以从 topic maps 中抽取有用的信息。topic maps 中确定了 3 个级别的信息，将其用作聚类的标准：(1) 分配给每个文档的主要主题 (major topics)，(2) 表示信息项的标签，比如 Country, City, Technology 等，以及与这些标签匹配的实际值，比如 Pakistan, Karachi 等。Xpath Queries 用于从 XTM 文件中抽取相关的主题、标签，以及标签所对应的实际值。使用上述的 3 个级别的信息可以生成 document-document 相似度矩阵。最后使用层次化凝聚式聚类算法得到聚类结果。