综合结构和内容的 XML 文档相似度计算方法

赵宁宁，梁意文

(武汉大学 计算机学院，湖北 武汉 430072)

摘 要：提出了一种综合考虑 XML 文档内容和结构信息的文档相似度计算方法.通过使用不同的方法分别计算文档内容信息相似度和结构信息相似度，然后赋予二者不同的权重将二者综合起来，得到文档的综合相似度.在真实数据集上的实验结果表明，综合结构和内容信息的方法能够提高计算 XML 文档相似度的准确性.

关键词： 内容相似度；结构相似度；XML 相似度；向量空间模型；路径频率

Combining Structure and Content Similaritiesmeasure for XML Document

ZHAO Ning-ning, LIANG Yi-wen

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract： This paper proposed a document similarity calculation method considering the XML document content and structure information in this paper. Different methods was used to calculate the document content similarity and structural information, and different emphasis was laied on them.Then the comprehensive similarity of the document can be attained. Experimental results on real data sets show that the method integrated structure and content information can improve the accuracy of calculation of XML documents similarity.

Key words： content similarity; structure similarity; XML similarity; VSM; path frequency

作者简介：

赵宁宁 女，（1992-），硕士研究生.研究方向为信息检索、人工免疫学.E-mail:349177535@qq.com.

梁意文 男，（1962-），教授.研究方向为人工免疫学.