频繁子树模式在中心词识别中的应用研究

田卫东，黄 勇
（合肥工业大学 计算机与信息学院，安徽 合肥 230009）

摘 要：中文问句中心词识别领域中，现有方法未能有效利用依存句法中的深层统计关系．为解决此问题并探究中心词在词的多维属性上的统计关系，首次提出多维树概念，给出多维频繁模式挖掘方案并应用于中文问句中心词识别中．针对此应用给出频繁子树模式精简及规则冲突解决方案，训练出一个中文中心词识别模型．此方法是典型的客观方法，实验表明，此方法有较好的稳定性、适应性与鲁棒性，且较条件随机场模型在准确率上有进一步提高．

关键词：条件随机场；依存关系树；频繁子树模式；模式精简；规则冲突；中心词
中图分类号：TP391 **文献标识码**：A **文章编号**：1000－7180(2015)11－0027－06

Study on the Application of Frequent Sub-tree Patterns in Focus Words Recognition

TIAN Wei-dong, HUANG Yong
(School of Computer & Information, Hefei University of Technology, Hefei 230009, China)

Abstract: In the field of Chinese Focus-words Recognition, current studies don't take full advantages of some deep statistical relationships in dependency syntax. To solve this problem and explore statistical relationships between Chinese focus words and the multiple properties of words, a new concept called Multi-Dimensional Tree (MDT) and a solution of mining frequent MDT pattern are proposed and applied. Solutions of condensing those frequent patterns and dealing with pattern conflicts are given, a Chinese focus words recognizer is trained. The method is a kind of typical objective method, the empirical results show that this method is good at stability, adaptability and robustness and can reach higher recognition accuracy rate than CRF model.

Key words: conditional random field; dependency relationship tree; frequent sub-tree pattern; condensing pattern; rule conflict; focus words

作者简介：
田卫东 男，(1970-)，副教授．研究方向为智能计算与数据挖掘．

黄 勇(通讯作者) 男，(1990-)，硕士研究生．研究方向为数据挖掘．E-mail：zhuixunxiyang@126.com

收稿日期：2015－01－23；**修回日期**：2015－03－12
基金项目：国家"八六三"高技术研究发展计划(2012AA011005)；国家自然科学基金(61273292)