计算机应用 2014, 34(1) 64-68 DOI: 10.11772/j.issn.1001-9081.2014.01.0064 ISSN: 1001-9081 CN: 51-1307/TP

先进计算

基于弱监督的属性关系抽取方法

杨宇飞,戴齐,贾真,尹红风

西南交通大学 信息科学与技术学院,成都 610031

摘要： 针对从中文百科中抽取属性关系时所面临的训练语料匮乏问题,提出一种利用极少人工参与的弱监督自动抽取方法。首先,利用中文百科条目信息模板中的半结构化属性关系回标条目文本自动获取训练语料;然后,根据朴素贝叶斯分类原理优化训练语料;最后,基于条件随机场(CRF)建立属性关系抽取模型。在互动百科中采集的数据集上进行实验,综合评价F值达到了80.9%。结果表明该方法能够获得质量较高的训练语料,并取得良好的抽取性能。

关键词： 关系抽取 弱监督 中文百科 朴素贝叶斯分类 条件随机场

Weakly supervised method for attribute relation extraction

YANG Yufei,DAI Qi,JIA Zhen,YI Hongfeng

School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China

Abstract: In order to solve the problem of insufficient training corpus for extracting attribute relation from Chinese encyclopedia, a weakly supervised method was proposed, which needed minimal human intervention. First, semi-structured attribute relations from Chinese encyclopedia entry infoboxes were used to tag entry texts for obtaining training corpus. Second, the optimized training corpus was obtained based on Naive Bayesian theory. Third, Conditional Random Field (CRF) was used to form attribute relation extraction model. The evaluation of F-score on the Hudong encyclopedia datasets was 80.9%. The experimental result shows that this method can enhance the quality of training corpus and runs a better extraction performance.

Keywords: relation extraction weak supervision Chinese encyclopedia Naive Bayes classification Conditional Random Field (CRF)

收稿日期 2013-07-29 修回日期 2013-09-12 网络版发布日期 2014-02-14

DOI: 10.11772/j.issn.1001-9081.2014.01.0064

基金项目:

国家自然科学基金资助项目;中央高校基本科研业务费专项资金资助项目;中国科学院自动化所复杂系统管理与控制重点实验室开放课题

通讯作者: 贾真

作者简介: 杨宇飞(1988-),男,河南驻马店人,硕士研究生,主要研究方向:信息抽取;戴齐(1963-),男,四川成都人,副教授,主要研究方向:数据挖掘、智能信息处理;贾真(1975-),女,河南开封人,讲师,硕士,主要研究方向:信息抽取,内容安全;尹红风(1964-),男,河南商丘人,教授,博士,主要研究方向:大数据、语义搜索。
作者Email: 729380204@qq.com

参考文献：

本刊中的类似文章

1．张毅 黄聪 罗元.基于改进朴素贝叶斯分类器的康复训练行为识别方法[J]. 计算机应用, 2013,33(11): 3187-3189
2．王科俊 吕卓纹 孙国振 阎涛.基于分层分数条件随机场的行为识别[J]. 计算机应用, 2013,33(04): 957-959
3．刘丹丹 彭成 钱龙华 周国栋.词汇语义信息对中文实体关系抽取影响的比较[J]. 计算机应用, 2012,32(08): 2238-2244
4．张微 汪西莉.基于超像素的条件随机场图像分类[J]. 计算机应用, 2012,32(05): 1272-1275
5．王希杰.词位标注汉语分词中上下文有效范围定量分析[J]. 计算机应用, 2012,32(05): 1340-1342
6．王健 冀明辉 林鸿飞 杨志豪.基于上下文环境和句法分析的蛋白质关系抽取[J]. 计算机应用, 2012,32(04): 1074-1077
7．钟军 田生伟 禹龙.Web文本中维吾尔语领域术语的自动发现[J]. 计算机应用, 2012,32(02): 407-410
8．刘磊 陈兴蜀 尹学渊 段意 吕昭.基于特征加权朴素贝叶斯算法的网络用户识别[J]. 计算机应用, 2011,31(12): 3268-3270
9．阳维 张树恒 王莲芸 张素.基于图像块分类器和条件随机场的显微图像分割[J]. 计算机应用, 2011,31(08): 2249-2252
10．张聪品 赵理莉 吴长茂.基于字词分类的层次分词方法研究[J]. 计算机应用, 2010,30(8): 2034-2037
11．应玉龙 李淼 乌达巴拉 朱海.基于条件随机场的蒙古语词性标注方法[J]. 计算机应用, 2010,30(8): 2038-2041
12．樊娜 蔡皖东 赵煜 李慧贤.中文文本情感主题句分析与提取研究[J]. 计算机应用, 2009,29(4): 1171-1173
13．何红洲 周明天.基于背景的个性化客户行为模型研究[J]. 计算机应用, 2009,29(12): 3283-3286
14．刘路 李弼程 张先飞.基于正反例训练的SVM命名实体关系抽取[J]. 计算机应用, 2008,28(6): 1444-1446
15．徐冰 郭绍忠 黄永忠.基于朴素贝叶斯分类算法的活跃网络结构挖掘[J]. 计算机应用, 2007,27(6): 1548-1550
16．蔡崇超 王士同.一种基于Bernoulli混合模型的不完整数据文本分类方法[J]. 计算机应用, 2007,27(5): 1235-1237