第一图书网, tushu007.com

<<词语计算与应用>>

图书基本信息

书名：<<词语计算与应用>>

13位ISBN编号：9787811354881

10位ISBN编号：7811354888

出版时间：2010-5

出版时间：暨南大学出版社

作者：刘华

页数：268

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

第一图书网, tushu007.com

前言

刘华博士的专著《词语计算与应用》（他谦称为一本“学术上摸索的小书”）就要出版了，希望我为他写一篇序，我答应了。
临近截稿日期了，我还没有开笔。
因为有一些急于处理的事情，而且又临近我住院的日子，所以我向刘华提出，要不那篇序就算了，有没有关系不大，不要耽误了书的出版。
刘华回复坚请，说还来得及，就是推迟几天出版，也要等老师的序来添色。
事情就可以从“添色”说起了。
新世纪的读者如果觉得刘华的“小书”《词语计算与应用》读起来有味道、实用，特别是文科的大学生、研究生，认为有新信息、新内容，那是此书本身所存在的“特色”，不是我所能“添”上去的。
正如刘华自己所说，他作为一个“计算语言学的门外汉”，经过几年在1和0的世界里纠结、挣扎，才获得了这些心得和成果。
不“纠结、挣扎”，一个“门外汉”怎么可能不仅进到门里，还登堂入室，拿到博士学位呢？
正所谓天道酬勤，一分耕耘，一分收获。
舒舒服服、投机取巧混文凭的人是有的，但这终究是自欺欺人，迟早会暴露。
刘华博士的努力是实在的、痛苦的、反复的，也是曲折向上的。
刘华自2002年起，在北京语言大学语言学及应用语言学博士点下攻读“语言信息处理”方向的博士学位。
作为一个文科出身的应用语言学的硕士，要以计算机为主要工具，以建设动态流通语料库为主要目标和研究手段，以语言信息处理为主要研究内容，对刘华来说，确实困难重重。

第一图书网, tushu007.com

<<词语计算与应用>>

内容概要

《词语计算与应用》共有四章，除了附录、后记外，核心内容词语的计算与应用，主要包括“领域新词语快速获取”、“词语分类和词语聚类”、“词语计算与辅助汉语教学”、“词语主题度计算与自动标引”几个方面，这些也都是目前理工科(包括图书馆的情报检索)关注的热门课题，属于人文学科与理工学科交叉的边缘领域。
语言信息处理、自然语言理解、人工智能、机器翻译等都是这一边缘领域的学科或课题。
理工专业人士研究此类项目时，要补充人文专业知识(如语言学)；人文专业人士研究此类项目，要补充理工专业知识(如计算机科学、数理科学)。
相对而言，补充人文专业知识较容易，补充理工专业知识则较困难。
也就是说，搞计算语言学，文科出身者比理工科出身者面临的压力大。
通常，理工科的人写的计算语言学的论著，满篇术语公式，文科读者觉得犹如读“天书”，但是刘华博士的《词语计算与应用》并非如此。
因为是文科出身的人写给文科出身的人读的书，作为一个“过来人”，他能设身处地为读者着想，每个术语都有诠释，甚至每个公式都有解读，文科的人读来并不觉得过于深奥晦涩。

第一图书网, tushu007.com

作者简介

刘华，男，1975年生，暨南大学副教授。
2005年毕业于北京语言大学中文信息处理专业，师从张普教授，获博士学位，主攻自动标引、计算语言学和计算语言学辅助汉语教学。
近五年来，在核心期刊发表论文二十余篇，多篇被EI索引；目前，主持国家级课题一项，省部级课题多项。

第一图书网, tushu007.com

<<词语计算与应用>>

书籍目录

第一图书网, tushu007.com

章节摘录

推而广之，我们还可以用此方法来自动发现词语的多个义项，并进行多义项的消歧。

2.3.4聚类词语集成　　2.3.4.1多类别映射　　我们最终完成了5万个种子词的词语聚类词表的自动构建。

由于聚类是在15大类中各自进行的，因此，有些种子词可能出现于多个大类中，并最终映射到具体的层级小类中。

例如，“交通”种子词，就属于“房产城市建设交通、汽车　汽车新闻、旅游黄金周、时政新闻　国内、时政新闻社会、经济消费理财消费生活、教育考试培训　职业技能　国家公务员考试、时政新闻　国际、科技　科普生活”等9个层级小类。

我们这一步的工作就是将种子词在多个类中的聚类词表中进行合成，当用户检索某种子词时，系统自动返回该种子词在不同类中的聚类词语表，而且根据种子词归属于各类的归属度将类由高到低排列。

例如，“接吻”种子词，按照其归属于各类的归属度，从高到低依次属于“生活男女两性迷情、时政新闻社会、教育性及教育、文艺艺术、时政新闻　国际、科技科普生活艾滋、旅游主题旅游蜜月旅游”，这一结果也和我们的语感基本一致。

种子词归属于各类的归属度是自动进行的，方法如下：　　如果种子词在几个类中都有，利用文本分类的向量空间模型算法计算种子词的特征向量和这几个类的特征向量之间的相似度，按照相似度从高到低排列即可。

文本分类的向量空间模型算法参见后文的介绍。

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com