

分类号: TP391 密级: _____
U D C: 681.3.02 学号: 9911107

学 位 论 文

基于文本层次模型的 Web 概念挖掘研究

——基于概念语义网络的自动标引和自动分类研究

章 成 志

指导教师姓名: 侯汉清 教 授
南京农业大学 南京 卫岗 210095

申请学位级别: 硕 士 专 业 名 称: 情报学 (挂靠农业经济与管理)

论文提交日期: 2002 年 6 月 论文答辩日期: 2002 年 6 月 日

学位授予单位: 南京农业大学 学位授予日期: 2002 年 月 日

答辩委员会主席: _____ 教 授

评 阅 人: _____ 教 授

南京农业大学

二〇〇二年六月

基于文本层次模型的 Web 概念挖掘研究

摘 要

本论文针对目前 Web 文本挖掘工具的不足之处,综合运用文献信息自动标引和自动分类技术、数据挖掘技术、模式识别技术、数据库技术、数理统计知识,构建了一个简单易行的信息提取模型,即文本层次模型,来实现因特网上三种结构类型数据的信息提取,其中包括了概念提取。本文的研究具有如下意义:使分类知识库建设系统化和流程化;提供因特网页面和普通文本的标引源选择方案及主题提取时的权重方案;进一步扩大同义词的识别能力;增强未登录词挖掘能力。

文本分类知识库的构建主要是利用了数据挖掘技术,数理统计知识,在进行关键词与分类号的相关度度量时,我们为了克服以前度量方法的缺陷,引入了 Dice 测度的方法。为了确定知识库的规模,我们 Web 概念挖掘系统的实际运行结果,进行抽样分析,选择了一个整体性能较好的分类知识库。

在进行 Web 文本的主题提取时,为了区分不同标引源的主题表达能力,本文根据一定规模的数据调查结果,确定了具有一定科学依据的权重方案,对文本不同标引源的测试获得了因特网页面和普通文本的标引源选择方案。

在同义词的识别上,首次引入《同义词词林》,作为语义体系,提出了基于《同义词词林》语义体系的同义词识别算法,并进行了词汇间的语义相似度度量,实现了多个大类中的同义词识别,提高同义词识别系统的识别能力。此外,在进行文本的自动分类时,将语义相似度代替了简单的关键词串和主题词串的匹配,提高了文本的自动分类能力。

为了解决未登录的挖掘问题,提出了基于字词正向扩展的未登录词识别方法,不同于 N-Gram 模型的是,本方法不需庞大的语料库,利用局部统计信息即可识别具有检索意义的未登录词。

本文最后给出了系统的实际测评结果,证明整个系统的可行性。

Web 概念挖掘系统采用 Borland Delphi6.0, Microsoft Visual C++6.0 以及 Microsoft Visual Foxpro6.0 开发。

关键词: 文本层次模型; Web 概念挖掘; Dice 测度;

同义词识别; 字词正向扩展; 未登录词

WEB CONCEPT MINING BASED ON TEXT LAYER MODEL

ABSTRACT

To improve the performance of Web text mining tools, this paper try on using automatic indexing and automatic classification techniques, data mining technology, pattern recognition technology, database technology and mathematical statistics knowledge to create a practical model, i.e. Text Layer Model (TLM),and it can extract information from three kinds of datum on the Internet. The significance of this paper is as follows: providing a new method to create the knowledge database used for classifying, providing the location weighting algorithm for information extraction, presenting a new methods to improve the performance of recognition of synonyms and unregistered words.

The creating of the knowledge database used for classifying is base on data mining technology and mathematical statistics knowledge. We use the Dice measure, support degree and confidence degree to create four kinds database of different dimensions through different thresholds of correlation degree and interesting degree. Lastly, we select one of database through the test by CWDS system.

To distinguishing the subject expression ability of different parts of text, including the Web pages, we have a investigative statistics and providing the location weighting algorithm for information extraction.

To enhance the ability of the recognition synonyms, we use the *synonyms dictionary* as the semantic system and providing the new algorithm of recognition synonyms base on the *synonyms dictionary*. We use this algorithm to calculate the similarity degree among the words and match the subject in the automatic classifying.

We provide a new method to enhance the ability of mining the unregistered words, i.e. recognition method base on the character or word expanding. Different from the N-Grams Modal, this method uses the location information of the text to recognize unregistered words.

At the end of the paper, we test and evaluate the CWDS system, RSS system and URWS system; the deficiency of systems is also detailed objectively.

KEYWORDS: web concept mining; text layer model; knowledge database;
recognition synonyms; unregistered words;

目 录

第 1 章 WEB 概念挖掘研究综述	1
1.1 WEB 数据挖掘研究综述	1
1.2 同义词识别研究综述	4
1.3 未登录词挖掘研究综述	7
1.4 本文的主要研究内容	9
第 2 章 文本层次模型的提出与建立	11
2.1 文本层次研究综述	11
2.2 文本层次模型的提出与建立	14
2.3 本章小结	19
第 3 章 WEB 概念挖掘系统的总体设计	21
3.1 系统概述	21
3.2 系统总体设计	22
3.3 试验数据概述	23
第 4 章 WEB 概念挖掘用分类知识库的制作	26
4.1 关键词(串)-分类号关联研究综述	26
4.2 关键词(串)-分类号关联方法	30
4.3 分类知识库的制作	37
4.4 分类知识库的性能测评	42
4.5 篇名知识库的制作	45
4.6 本章小结	47
第 5 章 WEB 概念挖掘中标引源权重方案的确定	49
5.1 标引源权重研究综述	49
5.2 标引源权重方案的确定	49
5.3 文本多主题挖掘	56
5.4 本章小结	58
第 6 章 基于语义体系的同义词识别	59
6.1 基于字面相似度和词素相似度算法的不足之处	59
6.2 《同义词词林》简介	60
6.3 基于《同义词词林》语义体系的相似度算法	60
6.4 同义词挖掘系统设计	68
第 7 章 WEB 概念挖掘中的未登录词挖掘	70
7.1 基于字词正向扩展的未登录词挖掘方法	70
7.2 未登录词挖掘系统设计	74

7.3 本章小结	74
第 8 章 WEB 概念挖掘系统的使用与测评	75
8.1 WEB 概念挖掘系统的使用与测评	75
8.2 同义词挖掘系统的使用与测评	80
8.3 未登录词挖掘系统的使用与测评	85
8.4 结束语	88
参考文献.....	90
致 谢.....	94