WWW 页面信息中特定内容的过滤研究*

胡熠 郑德权 赵铁军 于浩 王青松

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

E-mail: huyi@mtlab.hit.edu.cn

摘要: 伴随着 Internet 上日益丰富的信息资源，大量有害、不健康的中文信息渗入其中，严重影响人们正常的工作和生活，因此急需对 WWW 页面信息中的有害内容进行有效地识别。本文根据网络文本中有害内容的特点，提出了通过机器学习识别特定内容并过滤的新方法。该方法依据归纳学习策略，得到真实文本的有害度评测值，如果该评测值超过预定的阈值，则将其过滤。实验结果表明，根据学习方法开发的用于识别 WWW 页面中有害内容的实验系统，效果比较理想。

作者简介: 胡熠，男，1978 年生，硕士研究生，研究方向是机器翻译，计算语言学；郑德权，男，1968 年生，博士研究生，研究方向是计算语言学、机器翻译，自然语言处理；赵铁军，男，1962 年生，博士，教授，博士生导师，研究方向是机器翻译和计算语言学；于浩，男，1971 年生，博士，副教授，主要研究方向是自然语言理解和信息处理；王青松，男，1973 年生，硕士研究生，研究方向是计算语言学、自然语言处理。

关键词: 信息过滤，归纳学习，读音匹配，词性转移表

Research of Machine Learning Method for Specific Information Recognition on the Internet*

Hu Yi, Zheng Dequan, Zhao Tiejun, Yu Hao, Wang Qingsong

School of Computer Science & Technology

Harbin Institute of Technology(HIT), Harbin, 150001

E-mail: huyi@mtlab.hit.edu.cn

Abstract: With the available resources on the Internet becoming plentiful, a large amount of illegal harmful Chinese information is permeated among them .Therefore, some illegal text must be recognized and be filtered effectively in Web. Through analyzing a number of illegal harmful contents in information on the Internet, a new method is presented which recognizes specific information on the Internet by Machine Learning. For the more, the evaluation value-harmful extent-of whole real text will be obtained by inductive learning. If the evaluation value exceeds the preconcerted value, the real text will be filtered. The experiment proved that the effect of the experiment system developed on Machine Learning method was efficient for recognizing some pieces of illegal harmful Chinese information in Web.

Key words: Information Filtering; Machine Learning; Pronouncing Matching; Part of Speech Transfer-Form

* 本项目受到国家 863 资助（项目编号：2001AA114101）

* This project is supported by the National '863' High-Tech program of China (No. 2001AA114101)