P.O.Box 8718, Beijing 100080, China	Journal of Software March 2003,14(3):435-442
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2003 by The Editorial Department of Journal of Software

基于向量空间模型的文本过滤系统

黄萱菁, 夏迎炬, 吴立德

Full-Text PDF Submission Back

黄萱菁, 夏迎炬, 吴立德 (复旦大学 计算机科学与工程系,上海 200433)
第一作者: 黄萱菁(1972－),女,浙江平阳人,博士,副教授,主要研究领域为大规模文本信息处理.
联系人: 黄萱菁 Telephone: 86-21-65642192, E-mail: xjhuang@fudan.edu.cn
Received 2001-09-14; Accepted 2002-04-10

Abstract

Text filtering is the procedure of retrieving documents relevant to the requirements of specific users from a large-scale text data stream. First, the TREC (text retrieval conference) as well as its text filtering track are introduced, which is the most authoritative international evaluation conference on text retrieval, from the aspects of tasks, topics, corpus and evaluation metrics. Then a text filtering system based on vector space model is presented. This system is composed of two phases of training and adaptive filtering. During the training phase, feature selection and pseudo feedback are used to select the initial filtering profiles and thresholds. During the filtering phase, user feedback is utilized to modify the profiles and thresholds adaptively. This system took participate in the 9th Text Retrieval Conference in 2000, and ranked high among all the 15 systems from many countries. Good performance has been achieved, where the average precisions of adaptive and batch filtering are 26.5% and 31.7% respectively.

Huang XJ, Xia YJ, Wu LD. A text filtering system based on vector space model. *Journal of Software*, 2003,14(3):435~442.
http://www.jos.org.cn/1000-9825/14/435.htm

摘要

文本过滤是指从大量的文本数据流中寻找满足特定用户需求的文本的过程.首先从任务、测试主题、语料库和评测指标等方面介绍了文本检索领域最权威的国际评测会议——文本检索会议(TREC)及其中的文本过滤项目,然后详细地描述了基于向量空间模型的文本过滤系统.该系统由训练和自适应过滤两个阶段组成.在训练阶段,通过特征抽取和伪反馈建立初始的过滤模板,并设置初始阈值;在过滤阶段,则根据用户的反馈信息自适应地调整模板和阈值.该系统参加了2000年举行的第9次文本检索会议的评测,取得了很好的成绩,在来自多个国家的15个系统中名列前茅,其中自适应过滤和批过滤的平均准确率分别为26.5%和31.7%.

基金项目：Supported by the National Natural Science Foundation of China under Grant Nos.69873011, 69935010, 60103014 (国家自然科学基金); the National High Technology Development 863 Program of China under Grant No.863-306-ZD02-02-4 (国家863高科技发展计划); the National High-Tech Research and Development Plan of China under Grant No. 2001AA114120 (国家高技术研究发展计划); the Science and Technology Promotion Foundation of Shanghai of China under Grant No.995115005 (上海市科学技术发展基金); the Science and Technology Foundation of Fudan University of China (复旦大学科学技术基金)

References:

[1] Belkin N, Croft WB. Information filtering and information retrieval, two sides of the same coin. Communications of the ACM, 1992,33(12):29~38.

[2] Daniel EO. The Internet, Intranet, and the AI renaissance. Computer, 1997,30(1):71~78.

[3] David DL. The TREC-4 filtering track. In: Harman DK, ed. Proceeding of the 4th Text Retrieval Conference (TREC-4). Gaithersburg: NIST Special Publication, 1995. 165~180.

[4] Voorhees EM, Harman DK. Overview of the 9th text retrieval conference (TREC-9). In: Voorhees EM, Harman DK, eds. Proceedings of the 9th Text Retrieval Conference (TREC-9). Gaithersburg: NIST Special Publication, 2000. 1~14.

[5] Charles LW. Topic detection & tracking (TDT) overview & perspective. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. 1998. http://www.nist.gov/speech/publications/darpa98/index.htm, Lansdowne.

[6] Robertson S, Hull DA. The TREC-9 filtering track final report. In: Voorhees EM, Harman DK, eds. Proceedings of the 9th Text Retrieval Conference (TREC-9). Gaithersburg: NIST Special Publication, 2001. 25~40.

[7] Salton G. Developments in automatic text retrieval. Science, 1991,253(5023):974~979.

[8] Wu LD, Huang XJ. Large-Scale Chinese Text Processing. Shanghai: Fudan University Press, 1997. 102~118 (in Chinese).

[9] Buckley C, Salton G, Allan J. Automatic retrieval with locality information using SMART. In: Harman DK, ed. Proceedings of the 1st Text REtrieval Conference (TREC-1). Gaithersburg: NIST Special Publication, 1992. 59~72.

[10] Huang XJ, Wu LD, Ishizaki Hiroyuki, Xu GW. Language independent text categorization. Journal of Chinese Information Processing, 2000,14(6):1~7 (in Chinese with English Abstract).

附中文参考文献:
[8] 吴立德,黄萱菁.大规模中文文本处理.上海:复旦大学出版社,1997.102~118.

[10] 黄萱菁,吴立德,石崎洋之,徐国伟.独立于语种的文本分类方法.中文信息学报,2000,14(6):1~7.