一种高效的文本数据挖掘方法

文献类型：专利

作者 杨风雷；黎建辉；吴开超；薛正华；张波

发表日期 2011-11-28

专利号 CN102402606A

权利人 中国科学院计算机网络信息中心

中文摘要 本发明公开了一种高效的文本数据挖掘方法，属于信息技术领域。本方法为：1)文件预处理阶段将内容经分词后的原文件合并为若干新文件；2)数据映射阶段计算每一词语在新文件中的总频数、在其中每一原文件中的频数及相对频率pr等，并将结果发送到重定向模块中；3)重定向阶段计算每一Reduce任务的负载量payload，并为每一Reduce任务设置一负载指示器payi；4)判断当前词语是否已分配了Reduce任务；如果未分配，则将其分配给Reducej，且payj+pr*100≤payload成立；然后更新Reducej的负载指示payj；否则将当前词语分配给相应Reducei任务；5)数据规约阶段对分配的词语计算其最终的频数等参数；6)根据数据规约结果，提取设定范围内频数大于设定阈值的词语。本发明大大提高频数计算、数据挖掘效率。

公开日期 2012-04-04

申请日期 2011-11-28

专利申请号 201110385415.1

专利代理 北京君尚知识产权代理事务所(普通合伙) 11200

源URL [http://ircnic.ac.cn/handle/311056/1874]

专题 计算机网络信息中心_中国科学院计算机网络信息中心(2012年前)_专利

推荐引用方式 GB/T 7714 杨风雷,黎建辉,吴开超,等. 一种高效的文本数据挖掘方法. CN102402606A. 2011-11-28.

入库方式：OAI收割
来源：计算机网络信息中心

浏览	下载	收藏
8	0	0

其他版本

0931-8270076 发送邮件