(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104111971 B
(45)授权公告日 2018.03.13

(21)申请号 201410254061.0

(22)申请日 2014.06.09

(65)同一申请的已公布的文献号
申请公布号 CN 104111971 A

(43)申请公布日 2014.10.22

(73)专利权人 合肥工业大学
地址 230009 安徽省合肥市屯溪路193号

(72)发明人 任福继 刘宁 全昌勤 魏希权

(74)专利代理机构 安徽合肥华信知识产权代理有限公司 34112
代理人 余成俊

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 103324665 A,2013.09.25,
CN 103366017 A,2013.10.23,
CN 103279483 A,2013.09.04,
CN 103092921 A,2013.05.08,
CN 102663101 A,2012.09.12,
US 2014101293 A1,2014.04.10,
王琛等.一种改进的微博用户影响力评价算法.《信息工程大学学报》.2013,第14卷(第3期),第380-384页.

审查员 刘莹莹

权利要求书1页 说明书3页

(54)发明名称

过往微博数据收集与处理方法

(57)摘要

本发明公开了一种过往微博数据收集与处理方法,首先获取活跃微博用户ID,然后获取活跃微博用户微博数据,最后对微博数据进行处理。本发明改进了新浪第三方API,以弥补微博接口获得数据精确度的不足,能够满足过往微博数据收集与处理的要求。

CN 104111971 B

1.过往微博数据收集与处理方法，其特征在于：可以获得指定过往时间点或时间段内的微博数据；包括以下步骤：

(1)、获取活跃微博用户ID：

调用微博第三方API接口获取微博广场上公开的微博数据，公开的微博数据为微博作者的用户信息字段，其中包括用户UID、用户所在城市ID的信息；根据获取到的微博广场上公开的微博数据，提取出用户UID，去重后即为可用的活跃微博用户ID；

(2)、获取活跃微博用户微博数据：

将获取到的用户UID拆分为7个本地用户UID库，分别使用7个微博第三方API Token并行运行，提升单位时间内获取微博的数量；然后根据用户UID账号，调用微博第三方API应用接口获得对应账号下的所有微博数据文件，微博数据文件包括微博创建时间、微博信息内容、微博来源、微博作者的用户信息字段，微博数据文件保存为UTF-8格式的TXT文本文件，设微博数据文件为D；

(3)、微博数据处理：

根据相关热点事件，指定热点事件种子关键词，确定热点事件发生时间段；根据确定的热点事件时间段，从本地的微博数据文件D中提取指定事件时间段内的微博文本数据；微博文本数据包括微博创建时间、微博信息内容、用户昵称、用户所在地；提取后的微博精细内容文件本地保存为UTF-8格式的TXT文本文件，设微博精细内容文件为$\overline{D}$；根据用户所在地，对微博精细内容文件$\overline{D}$再次提取拆分为文本文件Da11以及文本文件类$D_{location_i}$，其中文本文件Da11为该热点事件对应的全国微博数据，文本文件类$D_{location_i}$为该热点事件对应的某城市微博数据，i≠0，为对应的城市代码；文本文件Da11以及文本文件类$D_{location_i}$中微博数据包括微博创建时间、微博信息内容，根据确定的热点事件发生时间段，进一步将文本文件Da11与文本文件类$D_{location_i}$拆分为该热点事件对应的全国微博数据单日数据集$D_{all\times day_t}$及该热点事件对应的某城市微博单日数据集$D_{location_i\times day_t}$，其中t为日期号。

过往微博数据收集与处理方法

技术领域

[0001] 本发明涉及微博数据处理方法领域，具体是一种过往微博数据收集与处理方法。

背景技术

[0002] 随着微博的兴起，这种包含了大量微观点并带有情感倾向的短文本迅速富集，微博文本分析成为热门研究方向。

[0003] 在微博数据搜集过程中，大量的微博数据搜集策略通常采用爬虫抓取方法，该方法抓取速度快、效率高，但是抓取的数据噪音大，虽然减少了数据搜集的时间，但是却成倍的增加了获得精确数据的预处理时间；且爬虫不稳定，常常面临被新浪封禁的危险。少量微博数据一般采用新浪微博第三方API进行调用搜集，该方法搜集的数据噪音少、区域明显，但是包含了大量的推送广告，又额外增加了无用数据比例。

[0004] 无论是爬虫方法还是传统的新浪第三方API调用，都无法大量获得指定域下的微博数据，特别是过往微博数据的处理，爬虫方法和新浪第三方API调用皆无法适用。

发明内容

[0005] 本发明的目的是提供一种过往微博数据收集与处理方法，以解决现有技术中爬虫方法或第三方API调用无法大量获取过往微博数据的问题。

[0006] 为了达到上述目的，本发明所采用的技术方案为：

[0007] 过往微博数据收集与处理方法，其特征在于：包括以下步骤：

[0008] (1)、获取活跃微博用户ID：

[0009] 调用微博第三方API接口获取微博广场上公开的微博数据，公开的微博数据为微博作者的用户信息字段，其中包括用户UID、用户所在城市ID信息；根据获取到的微博广场上公开的微博数据，提取出用户UID，去重后即为可用的活跃微博用户ID；

[0010] (2)、获取活跃微博用户微博数据：

[0011] 将获取到的用户UID拆分为7个本地用户UID库，分别使用7个微博第三方API Token并行运行，提升单位时间内获取微博的数量；然后根据用户UID账号，调用微博第三方API应用接口获得对应账号下的所有微博数据文件，微博数据文件包括微博创建时间、微博信息内容、微博来源、微博作者的用户信息字段，微博数据文件保存为UTF-8格式的TXT文本文件，设微博数据文件为D；

[0012] (3)、微博数据处理：

[0013] 根据相关热点事件，指定热点事件种子关键词，确定热点事件发生时间段；根据确定的热点事件时间段，从本地的微博数据文件D中提取指定事件时间段内的微博文本数据；微博文本数据包括微博创建时间、微博信息内容、用户昵称、用户所在地；提取后的微博精细内容文件本地保存为UTF-8格式的TXT文本文件，设微博精细内容文件为$\overline{D}$；根据用户所在地，对微博精细内容文件$\overline{D}$再次提取拆分为文本文件D_{all}以及文本文件类$D_{location_i}$，其

中文本文件Dall为该微博事件对应的全国微博数据，文本文件类$D_{location_i}$为该微博热点事件对应的某城市微博数据，i≠0，为对应的城市代码；文本文件Dall以及文本文件类$D_{location_i}$中微博数据包括微博创建时间、微博信息内容，根据确定的热点事件发生时间段，进一步将文本文件Dall与文本文件类$D_{location_i}$拆分为该热点事件对应的全国微博数据单日数据集$D_{all\times day_t}$及该热点事件对应的某城市微博单日数据集$D_{location_i\times day_t}$，其中t为日期号。

[0014] 本发明改进了新浪第三方API，采用并行多用户调用方式增加数据搜集流量；采用多信息点覆盖搜集微博数据，以弥补微博接口获得数据精确度的不足，能够满足过往微博数据收集与处理的要求。

具体实施方式

[0015] 过往微博数据收集与处理方法，过往微博数据是指用户在当前时间以前所发布的微博数据，其特点是数据固定，事后分析方便，包括以下步骤：

[0016] (1)、获取活跃微博用户ID：

[0017] 调用微博第三方API接口获取微博广场上公开的微博数据，公开的微博数据为微博作者的用户信息字段，其中包括用户UID、用户所在城市ID信息；根据获取到的微博广场上公开的微博数据，提取出用户UID，去重后即为可用的活跃微博用户ID；

[0018] (2)、获取活跃微博用户微博数据：

[0019] 将获取到的用户UID拆分为7个本地用户UID库，分别使用7个微博第三方API Token并行运行，提升单位时间内获取微博的数量；然后根据用户UID账号，调用微博第三方API应用接口获得对应账号下的所有微博数据文件，微博数据文件包括微博创建时间、微博信息内容、微博来源、微博作者的用户信息字段，微博数据文件保存为UTF-8格式的TXT文本文件，设微博数据文件为D；

[0020] (3)、微博数据处理：

[0021] 根据相关热点事件，指定热点事件种子关键词，确定热点事件发生时间段；根据确定的热点事件时间段，从本地的微博数据文件D中提取指定事件时间段内的微博文本数据；微博文本数据包括微博创建时间、微博信息内容、用户昵称、用户所在地；提取后的微博精细内容文件本地保存为UTF-8格式的TXT文本文件，设微博精细内容文件为$\overline{D}$；根据用户所在地，对微博精细内容文件$\overline{D}$再次提取拆分为文本文件Dall以及文本文件类$D_{location_i}$，其中文本文件Dall为该微博事件对应的全国微博数据，文本文件类$D_{location_i}$为该微博热点事件对应的某城市微博数据，i≠0，为对应的城市代码；文本文件Dall以及文本文件类$D_{location_i}$中微博数据包括微博创建时间、微博信息内容，根据确定的热点事件发生时间段，进一步将文本文件Dall与文本文件类$D_{location_i}$拆分为该热点事件对应的全国微博数据单

日数据集$D_{all\times day_t}$及该热点事件对应的某城市微博单日数据集$D_{location_i\times day_t}$，其中t为日期号。