基于海量信息过滤的微博热词抽取方法①

汪 洋[1], 帅建梅[1], 陈志刚[2]

[1](中国科学技术大学 信息科学技术学院, 合肥 230027)
[2](安徽科大讯飞信息科技股份有限公司 研究院, 合肥 230088)

摘 要: 针对海量微博信息, 提出一种多步骤的热词抽取方法. 首先, 选择用户行为特性、微博信息的文本特征构建用户行为模型, 并在此基础上提出一种基于规则的话题树生成过滤算法, 筛除了微博中大量无关信息, 进而对生成的话题树修剪优化; 然后, 根据话题树的节点内容, 使用词频及其波动特性设计热词抽取算法, 获取微博的热词信息. 实验数据表明, 该方法能大大减小输入的数据规模, 同时较好的保留重要信息, 有效实现热词抽取.

关键词: 中文微博; 用户行为模型; 海量信息过滤; 热词抽取; 幂律分布

Hot Word Extraction for Microblog Based on Massive Data Filtering

WANG Yang[1], SHUAI Jian-Mei[1], CHEN Zhi-Gang[2]

[1](School of Information Science and Technology, Science and Technology of China, Hefei 230027, China)
[2](iFLYTEK Research, Hefei 230088 China)

Abstract: This paper presents a Chinese microblog hot words extraction algorithm based on massive data Filtering. Firstly, it chooses the user behaviour characteristics and text characteristics to create user behavior models, and filters massive data to create topic-trees by a fast algorithm based on rules. Then, it uses hot words extraction algorithm to get the hot topic of topic-trees by word frequency feature. The experiment results show that the proposed algorithm can reduce the scale of the input data, with keeping lots of important information to extract hot words.

Key words: Chinese microblog; user behavior models; massive data filtering; hot word extraction; power law distribution

1 概述

微博(microblog)正在成为互联网中越来越重要的信息交流平台, 以新浪微博为例, 根据新浪 2011 年第四季度财报, 其注册用户已经突破 3 亿大关, 用户每日发布信息量超过 1 亿条. 微博在大量热点事件中扮演了传统媒体所不具有的信息快速发布的传播的角色, 同时, 微博平台可以在极短时间内汇聚相当数量的用户对同一热门事件的讨论信息, 如 2011 年的“7.23 温州动车追尾事故”、“郭美美事件”、“药家鑫事件”都是在微博平台首先发布并获得大量用户的迅速关注. 这些特点是其它传媒平台所难以企及的. 从微博信息抽取热门词语可以了解微博信息动态, 掌握舆论动向.

微博作为一种新型的网络媒体, 较之门户网站、博客等其它网络平台有着自己鲜明的特点. 首先, 微博的信息长度被限制在较小范围内(一般不超过 140 字), 使得每条微博包含的信息量相对一篇新闻或博客大大减少; 其次, 发布微博的门槛被大大降低, 任何人都可以发表内容而不必具有专业的文学撰写水平; 同时, 微博具有独特的用户交互功能, 一个用户可以关注自己兴趣的用户, 成为其“粉丝”, 同时也会被其它用户关注, 拥有自己的“粉丝”. 一个用户发表的微博可以被成百上千的“粉丝”同时阅读, 也可以同时关注大量的微博信息, 大量用户之间构成了复杂的虚拟社交网络, 这种虚拟社交网络类似现实中的社交网络, 但其中信息的传播速度更为快速便捷, 数以千万计的用户可以同时交流信息、发表评论, 只要信息能吸引用户, 就能在极短的时间内快速传播到极为广泛的范围.

由于微博具有的以上特性, 使得如何从海量的、包

① 收稿时间:2012-03-13;收到修改稿时间:2012-04-18

含大量无用信息的数据集中获取真正有用的热词信息成为巨大的挑战. 国内外在相关领域做了大量研究, 文献[1]提出人类动态行为并非符合传统的泊松分布, 而是非均匀的呈现出短期的活跃性及长期的休眠性, 相邻两个任务时间分布呈现“长尾”特性, 符合幂律分布. 此后, 许多研究者也证明许多人类活动符合幂律分布的特点, 例如: 网页浏览[2]、网络视频播放[3]、社交网络[4]等. 文献[5]研究了英文微博twitter中的信息分布情况, 文献[6]研究了微博用户之间的“追随”关系, 文献[7]研究了微博信息扩散的模型, 文献[8]考虑了微博联系人的关联关系和文本关联关系, 进而发现微博的热门话题. 上述研究大多基于英文微博信息及西方用户习惯构建模型, 针对文化差异是否会导致中文用户行为模型的不同, 还有待研究. 此外, 部分文献的算法研究仅基于小数据样本空间, 面对海量数据时, 复杂度较高的算法带来的时间空间开销可能难以接受.

本文基于以上理论基础, 考虑处理海量信息时算法复杂度的限制, 将热词抽取分为两个步骤: 首先, 考察多种用户行为特性及微博信息的文本特征, 构建用户行为模型, 提出一种基于规则的话题树生成过滤算法, 从而筛除大量无关信息; 然后, 利用词频及其波动特征设计热词抽取算法, 获取微博热词信息. 实验证明, 过滤算法在大大减小数据规模的同时能够较好的保留重要信息, 实现热词抽取.

2 微博海量信息的过滤及热词抽取

微博热词抽取主要包含四个主要工作: 微博数据收集及预处理、微博用户的行为模型的分析、设计信息过滤算法、设计热词抽取算法.

2.1 微博数据收集及预处理

微博平台提供了专用的 API 编程接口, 便于相关软件从微博平台获取信息, 本文借助大量微博 API 代理, 通过浏览微博用 用户的好友、以及其好友的好友的消息进行抓取, 每天循环获取一次, 从而累积大量数据集, 微博平台选择新浪微博与腾讯微博.

根据对新浪微博、腾讯微博分析, 转发模式以“//@用户名:内容”或“||@用户名:内容”为主, 以此为规则将信息按照转发规则分割为多个“用户/信息”. 例如: 输入一条信息“//@A:msgA //@B:msgB”, 输出“用户/信息”集合: { “A:msgA”, “B:msgB” }.

按照以上描述, 设计分割算法如下:

步骤 1: 输入“用户名/信息”, 在“信息”中查找@符号, 若找到, 转步骤 2. 否则, 保存“用户名/信息”到结果中, 算法结束.

步骤 2: 匹配规则“//@用户名:信息”,将匹配到的结果格式化为“用户名/信息”, 递归调用分割算法.

步骤 3: 删除信息内匹配部分, 转步骤 1.

现实数据中, 由于微博系统并没有强制用户使用上述的转发格式, 使得实际中可能存在多种情况, 例如, 转发的信息可能在“//”与“@”间存在多个空格, 半角“:”可能被全角“:“代替等等, 需要编写多种正则表达式匹配模式.

2.2 微博用户的行为模型的分析

利用 2.1 节获取的数据对比分析, 选择表 1 中列举的统计特征, 考察特征之间的关系建立模型.

表 1 选择的统计特征

特征名称	特征含义
信息长度(msg_length)	信息中包含字符个数
信息比率(msg_count_rate)	满足条件的信息个数占信息总数的百分比。
特殊符号数(symbol)	信息中包含的超链接或表情中包含的字符个数。
用户发布信息数(usermsg_count)	单个用户在单个数据集中发布的信息总数。
用户比率(user_count_rate)	满足条件的用户数占用户总数的百分比。
用户被转发次数(user_rt_count)	单个用户在单个数据集中发布的全部信息被转发的总数。
信息转发次数(msg_rt_count)	单个信息在单个数据集中被转发的总数。

2.2.1 信息长度-信息比率

由图 1 可以分析得出以下关系:

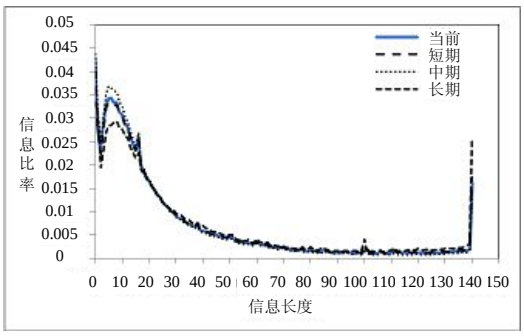

图 1 信息长度-信息比率关系

(1)信息长度与信息比率在当 15≤msg_length<130

时，基本满足幂律分布关系，即模型公式为 msg_count_rate =c*msg_length^{-r}，取四组数据平均值计算可得 c=1.495, r=1.50.

(2)msg_length=0，信息中不包含用户的原创信息，用户完全转发别人的信息，此类信息比率稳定在 4%左右.

(3) msg_length>130，信息比率反而增长，尤其是 msg_length=140 时，信息比率达到 1.8%左右，表明有一部分用户存在尽可能达到字数上限的倾向.

(4) 1≤msg_length＜15，此类信息大多较为口语化，分布与 15≤msg_length 的包含有意义内容的信息不同，比率稳定在 2%-3%左右.

2.2.2 特殊符号-信息比率

微博具有口语化的特点，信息中存在大量非文字符号(如特殊的表情符号、超级链接)与文字混用的情况. 信息中的非文字符号含有的信息往往对提取热门主题没有帮助，由表 2 可知，超链接及表情符号在微博信息中占有一定比重，且分布较稳定，随时间变化不大，可以完全过滤，减少一定运算存储代价.

表 2 特殊符号-信息比率关系

符号	当前	短期	中期	长期
超链接	0.127	0.131	0.107	0.168
表情	0.225	0.219	0.228	0.172

2.2.3 用户发布信息数-用户比率

由图 3 分析得出以下关系:

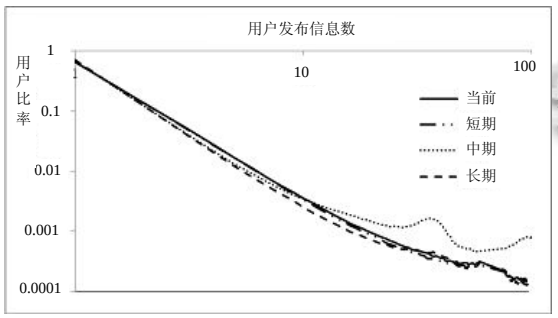

图 3 用户发布信息数-用户比率关系

(1)对于“当前”、“短期”、“长期”数据集，用户发布信息数与用户比率满足幂律分布，user_count_rate =c*usermsg_count^{-r}，取三组数据平均值得 c=0.533, r=2.17.

(2)“中期”数据比较其它时期数据，usermsg_count <10 时 user_count_rate 明显减少，usermsg_count>10 时 user_count_rate 明显增加，说明在春节假期中，用户更倾向于发较多的信息.

综合(1)(2)可知，用户发布信息数在普通时期相对比较稳定，在特殊时间(如春节等假期)会有一定的波动变化.

2.2.4 用户被转发次数-用户比率

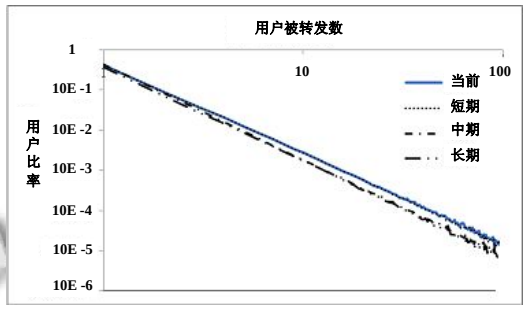

图 4 用户被转发次数-用户比率

表 3 未被转发的用户比率

数据集	用户比率
当前	0.393343
短期	0.391291
中期	0.402425
长期	0.490581
平均值	0.41941

由表 3、图 4 分析得出以下关系:

(1)当 user_rt_count>0 时，用户被转发次数与用户比率满足幂律分布, user_count_rate = c*user_rt_count^{-r}, 取四组数据平均值计算得 c= 0.441, r=2.30.

(2)user_rt_count=0，user_count_rate 稳定在 40%左右.

综合(1)(2)可知，有大量的用户被转发信息数很少，基本会不受到关注；有少量的用户发布的信息会受到广泛关注，进而形成热点话题，对这一部分用户要格外关注.

2.2.5 信息转发次数-信息比率

由表 4、图 5 分析以下关系: (1)当 msg_rt_count>0, 信息转发次数与信息比率满足幂律分布，msg_count_rate = c* msg_rt_count^{-r}，取四组数据平均值计算得 c=0.061, r=2.61. (2) msg_rt_count=0, msg_count_rate 稳定在 89%左右.

综合(1)(2)可知，大量信息被转发数很少，这些信息往往带有极强的个人色彩，不会受到广泛关注；有

少量的信息被大量转发, 进而形成热点话题. 可以考虑过滤关注度较低的信息, 从而提高热门话题在语料中的比重.

表 4　未被转发的微博信息比率

当前	短期	中期	长期
0.899621	0.902723	0.879866	0.911595

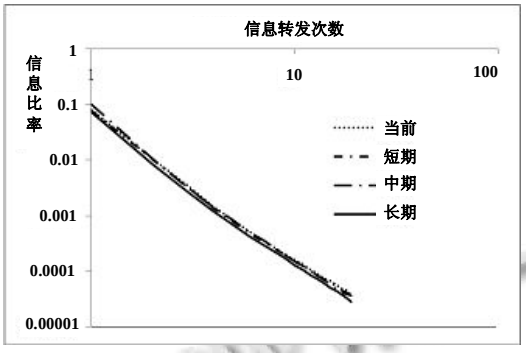

图 5　信息转发次数-信息比率关系

2.3 基于用户的行为模型的信息过滤

微博中信息的传播方式是由一个人发表信息开始, 围绕该信息, 许多人通过转发及评论联系在一起, 构成树状结构关系, 这种结构被称为话题树, 如图 6 所示. 微博信息是由大量话题树构成的, 每一颗话题树中的信息往往都是与根结点主题相关, 所以, 通过构建并分析话题树, 可以获取该类中信息的焦点热词.

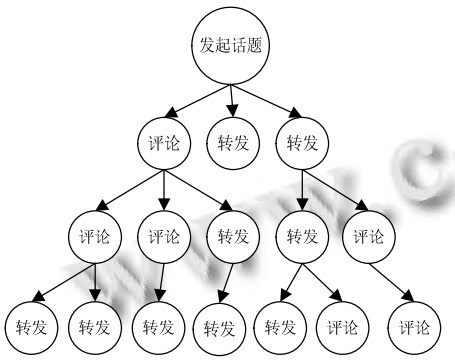

图 6　话题树结构图

2.3.1 话题树生成算法

根据 2.1 节的信息分割算法可以将一条包含了转发评论内容的信息分割为多个“原创信息”的形式, 并且这些“原创信息”构成了一条转发链式关系, 以此构建话题树. 由于微博中存在的信息量极为巨大, 利用全部信息构建话题树需要消耗大量的时间空间, 同时根据 2.2 中的数据分析, 大量的数据都是无关信息, 可以通过一定的规则将此过滤. 故选择信息长度(msg_length)、特殊符号(symbol)、用户发布信息数(usermsg_count)、用户被转发次数(user_rt_count)、信息转发次数(msg_rt_count)等特征, 制定以下判定规则, 判定规则中的各个阈值由 2.2 中构建的模型公式计算, 并通过实验确定.

(1)特殊符号过滤: 信息中包含表情、超链接等特殊符号时, 删除该特殊符号

(2) 信息长度判定: msg_length > MIN_MSG_LEN 时, 保留该信息, 否则该信息不能作为信息转发链的起点.

(3) 用户判定: usermsg_coun) > MIN_USERMSG_COUNT, 或 user_rt_count > MIN_USER_RT_COUNT, 保留该用户转发的信息, 否则过滤该信息.

(4) 转发判定: msg_rt_count > MIN_MSG_RT_COUNT 时, 保留该信息, 否则过滤该信息.

结合以上规则, 输入微博信息, 输出过滤生成的话题树集合, 算法流程如图 7 所示.

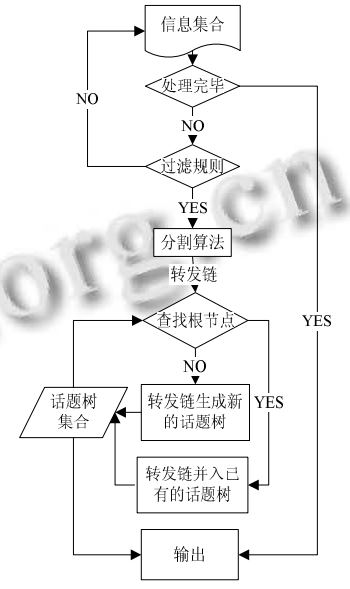

图 7　话题树生成算法

2.3.2 话题树修剪算法

由 2.3.1 生成话题树集合, 在同一颗话题树中, 主要围绕着一个话题讨论. 由 2.2 节中获得的用户行为模型可知, 话题树中大量节点不会出现与话题相关的关键词, 而是诸如“转发”、“支持”、“反对”等表达用户态度或仅仅是聊天的内容. 同时, 话题树的高度

和宽度代表了话题扩散的范围, 若话题的扩散范围很小, 说明该话题被关注的程度很低, 该话题往往具有很强的个人色彩或仅仅代表极小范围群体的喜好, 不能用于代表大范围内话题的特点, 此类话题树应该被删除. 综合考虑以上多种因素, 设计话题树的修建算法如下, 其中阈值 TREE_DEPTH、TREE_WIDTH、MIN_MSG_LEN 来自实验数据.

步骤 1: 输入话题树 T, 若 depth(T) < TREE_DEPTH 或 width(T) < TREE_WIDTH, 删除话题树, 算法结束. 否则, 继续下一步骤.

步骤 2: 遍历 T 的节点, 若节点 N 的 msg_length < MIN_MSG_LEN, 则使用根节点的内容替换该节点.

2.4 热词抽取算法

经过 2.3 节过滤数据获得话题树集合, 将其中全部节点内容按照文本输出, 经过中文分词可以求出每个词语的频度(WF, word frequency), 定义每天数据集中词语的热度(HT, hot words)为公式(1), 将频度平滑后的结果. 若 WF=0, 令 HT= - 10.

$$HT=\log(10+\log(WF)) \quad (1)$$

定义词语的热度增幅为公式(2)所示, today_ht 是当天词语的平均热度, short_ht 是最近短期(如 7 天内)的词语的平均热度, long_ht 是最近长期(如 30 天)的词语的平均热度, 公式中强调以词语在较长时间内的平均状态为主, 同时兼顾考虑突发事件带来的词语平度变化. 系数及数据集选择来自经验值.

$$Rise\_ht=today\_ht/short\_ht*0.2+today\_ht/long\_ht*0.8 \quad (2)$$

定义词语的最终热度为公式(3)所示, 公式综合考虑词语的当天热度及热度波动状况, 能反映出某些长期热门的词语及短期突发事件出现的词语.

$$Finht=0.9*today\_ht+0.1*\ Rise\_ht \quad (3)$$

按照词语的词语的最终热度为排序, 同时提出某些长时间存在的常见话题, 例如: “生日快乐”、“微博应用”等等, 最终获得热词排行榜.

3 实验及结果分析

实验选择从新浪微博、腾讯微博抓取的 2012 年 2 月 21 日数据作为测试集合, 总计包含 16331694 条信息, 涉及 2641075 名用户. 实验中根据大量经验数据, 设置参数 MIN_MSG_LEN = 4, MIN_USERMSG_COUNT=1,MIN_USER_RT_COUNT=4, MIN_MSG_RT_COUNT = 4, TREE_DEPTH = 4、TREE_WIDTH = 4. 从定性与定量两个方面分析实验结果.

3.1 定性分析

抽取获得的热词前 10 名及得分如表 5 所示, 从表中可以看出, 本文算法获取的热词在一定程度上可以反映当天微博信息中的热门词语, 例如: “个税”、“活熊取胆”、“篮网 vs 尼克斯”. 但也有一些垃圾字串不能包含, 如: “打开手机”、“南宁”、“响当当”等, 需要考虑使用停用词表等方法将垃圾字串过滤, 进一步优化结果.

表 5 top10 的热词

排名	热 词	得 分
1	穿越	0.91513
2	个税	0.870149
3	早高峰	0.848148
4	活熊取胆	0.843371
5	打开手机	0.796476
6	南宁	0.777365
7	吴奇隆	0.772218
8	响当当	0.767461
9	篮网 vs 尼克斯	0.766932
10	2012	0.763467

3.2 定量分析

3.2.1 过滤算法对信息长度分布的影响

图 7 中显示经过过滤前后信息长度-信息比率分布变化, 可以看出, 过滤前大量信息长度集中在 1≦msg_length≦20 个字长的区间. 过滤后删除了 msg_length <4 的信息, 同时大大提高了 msg_length 在 10 左右及 msg_length 在 30 左右的信息比率, 此外 msg_length>40 的长信息分布影响不大. 对由上文构建

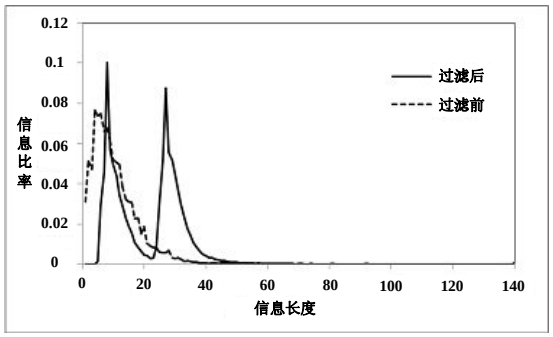

图 7 信息长度-信息比率分布图(过滤前后)

的用户行为模型可知, msg_length 在 10 左右的信息代表

了大量口语化的信息特征, msg_length 在 30 左右及 msg_length>40 包含了较多非个人化的信息特征. 所以, 经过过滤的信息很大程度上能代表原信息的信息分布特征.

3.2.2 过滤算法对词频分布的影响

表 6 统计了过滤前后微博信息中词语频度分布的变化, 分析获得以下结论:

表 6 过滤前后词频分布变化

log(词频)区间	过滤前词语数	过滤后词语数	相同词语比率	减少词语数
≧10e-2	5	5	100%	0
≧10e-3	90	115	100%	-25
≧10e-4	1241	1303	99.11%	-62
≧10e-5	6898	5669	93.95%	1229
≧10e-6	17669	12463	78.23%	5206
≧10e-7	25036	16378	60.75%	8658
≧10e-99	32342	16379	47.24%	15963

(1)过滤后词语数为 16379, 较过滤前的 32342 减少了 15963 个, 接近 50%左右.

(2)删除的词语中主要为低频词语, 过滤前在 log(词频)< 10e-5 的区间中有 25444 个词语, 过滤后仅有 10710 个词语, 减少的词语占总词语总删除量的 92.3%, 尤其是 log(词频)< 10e-7 的极低频区间中, 过滤前有 7306 个词语, 过滤后仅有 1 个词语. 删除低频词语可以减少待处理的数据规模, 提高算法效率.

(3)在 log(词频) ≧10e-5 的较高频区间, 过滤前与过滤后的词语数变化不大, 相同词语比率达到了 93.95%, 这样对整体热词频度计算影响不大. 尤其由于删除了大量低频无关词语, 从而提高了少量热门词语的频度, 所以在 log(词频) ≧10e-4 的区间内的词语数反而有所增加, 改善了热词抽取算法的结果.

综合 3.1 与 3.2, 从定性及定量两个方面证实本文算法能够较好的过滤海量微博信息, 在筛除大量无关信息的同时, 保留了大量重要信息, 从中有效获取了当前热词.

4 结语

微博信息存在着数据量巨大、内容形式多变、存在大量未登录词等特点, 本文提出的过滤及热词抽取算法能在一定程度上过滤海量信息, 并抽取微博热词, 具有速度快, 算法开销小的特点. 在未来工作中, 可以考虑引入未登录词发现的方法获取网络新词, 使用语义理解的方法挖掘微博信息, 更深入的发觉用户间信息传播模型等. 在大规模实时抽取中, 还可以采取分布式的方法提高算法效率.

参考文献

1 Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature,2005,435:207.

2 Dezs Z, Almaas E, Lukács A, et al. Dynamics of information access on the web. Phys Rev E, 2006,73(6):066132.

3 Zhou T, Kiet HAT, Kim BJ, et al. Role of activity in human dynamics. Europhys Lett, 2008,82(2):28002.

4 Grabowski A, Kruszewska N, Kosiński RA. Dynamic phenolmena and human activity in an artificial society. Phys Rev E, 2008,78(6):066110.

5 曹鹏,李静远,满彤.Twitter 中近似重复消息的判定方法研究.中文信息学报,2011,25(1):20–27.

6 刘志明,刘鲁.微博网络舆情中的意见领袖识别及分析.系统工程,2011,29(6):8–16.

7 许晓东,肖银涛,朱士瑞.微博社区的谣言传播仿真研究.计算机工程,2011,37(10):272–274.

8 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795–1802.

(上接第 110 页)

参考文献

1 何小明,蒋永东.气象信息共享服务业务平台用户认证系统的设计.现代电子技术,2007,21(2):12–16.

2 李集明,沈文海,王国复.气象信息共享平台及其关键技术研究.应用气象学报,2006,17(5):34–36.

3 琚玲,赵芳.ORACLE 数据库连接配置浅析及故障排除.气象科技,2009,37(4):448–451.

4 何险峰,蒋丽娟,等.公共气象服务网站数据的及时发布.气象科技,2011,(4):483–488.

5 夏正龙,尹新怀,欧阳计跃等.SharpMap 在实现湖南省降水色斑图显示中的应用.计算机系统应用,2011,20(2):134–136.

6 苏厚雄,王婉茹.基于 Ajax 和 PHP 数据分页的实现.计算机系统应用,2012,21(2):218–220.