基于相关反馈的微博相似主题时序查询

包红云 李秋丹 宋双永 高 珩
(中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190)

摘 要 提出了一种基于相关反馈的微博相似主题时序查询方法。该方法通过考虑用户对不同查询结果是否满意的反馈情况,建立修改度量系数的目标函数,从而实现微博中体现用户兴趣的主题时序相似性计算,为用户提供更满意的相似主题时序查询结果。基于该方法设计了一个可视化的微博相似主题时序查询系统,在微博代表性网站-Twitter数据集上进行的实验,表明了该方法在微博背景下的相似主题时序查询中的有效性。

关键词 微博客,主题时序,相似查询,相关反馈
中图法分类号 TP391 **文献标识码** A

Relevance Feedback-based Search of Topic Time Series Similarity in Micro-blogging

BAO Hong-yun LI Qiu-dan SONG Shuang-yong GAO Heng
(State Key Laboratory of Management and Control for Complex Systems Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract A new approach based on relevance feedback was proposed for the topic time series similarity search in micro-blogging. By considering whether the user is satisfied with the returned time series, we established an objective function for learning the coefficient of the unique metric, which reflects the user's accurate interest. Therefore, the approach can provide the user with more satisfying topic time series in micro-blogging. We also developed a topic time series similarity search system in micro-blogging based on the new approach. Experiment results on Twitter data show the effectiveness of our proposed approach.

Keywords Micro-blogging, Topic time series, Similarity search, Relevance feedback

1 引言

微博是一个为用户提供信息分享、传播及获取服务的新型网络平台,以其交互便捷、信息传播快速等特点,吸引了大量用户的积极参与[1]。越来越多的用户将微博作为了解网络热点主题相关信息的重要社会媒体平台。在微博中,热点主题出现后,用户通常以发帖的方式围绕该主题展开讨论。单位时间内用户发帖的数量能够体现用户群体在相应时间对特定主题的关注程度,相应地,用户对主题的关注度随时间的变化情况则能够很好地反映主题的发展状况。微博中主题关注度时序分析是新兴的研究热点[2]。

微博中相似主题时序查询能够基于用户特定时间段内感兴趣的主题时序检测出与用户感兴趣主题时序相似的其他主题时序,帮助用户更好地了解查询主题的整体发展状况,并且为用户预测主题未来发展趋势提供参考依据[3]。为了更好地为微博用户提供其感兴趣的相似主题时序,用户可对系统返回的主题时序根据其兴趣进行进一步反馈。目前将反馈技术引入到时序相似分析领域的研究主要针对心电图时序、电力负荷时序和美国失业率时序,基于用户的正反馈研究改进了查询时序[4-6]。微博中主题更新速度快,主题时序每个点对应的用户关注度具有表征用户群体兴趣走向和变化迅速的特点。目前,在微博中还未出现基于相关反馈进行相似时序查询的相关研究。另外,反馈信息中通常包括用户喜欢的时序和不喜欢的时序两个方面,它们从不同的角度反映了用户兴趣,因此,本文基于微博主题关注度时序的特点,研究综合考虑用户正、负反馈的时序相似查询方法,以为用户提供更满意的相似主题时序。基于查询获取的相似主题时序,用户可通过观察分析已有时序的发展变化情况,预测其感兴趣主题的未来发展趋势,从而实现对此主题的有效监管。

基于以上分析,我们提出了一种基于相关反馈的微博相似主题时序查询方法。对于指定时间段的查询主题,该方法首先统计出其对应的用户关注度时间序列,然后计算主题时序间的相似性,并根据与查询时序相似性的大小进行排序,为用户提供相似时序查询结果。其次,针对微博主题时序特点,设计了一种约束用户正反馈时序逼近查询时序且负反馈时序远离查询时序的目标函数,并基于该函数学习度量权重,进而实现了体现用户兴趣偏好的度量设计。最后,根据修改后的度量函数检索相似主题时序,为用户提供更满意的检索结果。基于所提方法,本文实现了一个可视化的微博相似主题时序查询原型系统,并通过实验验证了该系统在微博相似主题时

到稿日期:2012-06-15 返修日期:2012-10-10 本文受国家自然科学基金(61172106),北京市自然科学基金(4112062)资助。
包红云(1985—),女,博士生,主要研究方向为信息检索等,E-mail:hongyun.bao@ia.ac.cn;李秋丹(1976—),女,博士,副研究员,主要研究方向为信息检索、数据挖掘、移动电子商务等;宋双永(1986—),男,博士生,主要研究方向为信息检索等;高 珩(1987—),男,硕士生,主要研究方向为信息检索等。

序查询中的有效性。本文第 2 节介绍基于相关反馈的微博相似主题时序查询的方法；第 3 节描述本文设计的原型系统、实验数据、实验结果及其分析；最后总结全文。

2 基于相关反馈的微博相似主题时序查询方法

本文提出的基于相关反馈的微博相似主题时序查询方法的流程如图 1 所示。该方法主要包含如下 3 个步骤。1）主题时序表示，即对于给定的查询主题以及已有的主题数据集，统计其对应的用户关注度时间序列；2）主题时序相似性计算，即利用归一化的滑动时间窗欧氏距离计算查询主题时序与数据集中主题时序的相似性，根据相似值的大小进行排序，为用户提供与其感兴趣主题时序相似的主题列表；3）基于相关反馈的相似性度量建立，即用户可根据其兴趣对上述提供的相似主题时序给予相关与否的判断，根据反馈信息学习建立相似度量，进而为用户提供更为满意的相似主题时序列表。

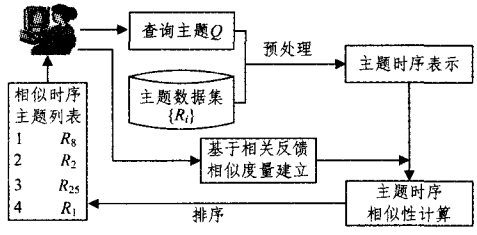

图 1 基于相关反馈的微博相似主题时序查询方法流程图

2.1 微博中主题时序的表示

微博中，用户针对热点主题发帖是表达对该主题关注的一种重要方式。单位时间内，用户围绕主题发表的帖子数量反映了用户群体在对应时间内对该主题的关注程度。相应地，随着时间推移，主题用户关注度的变化则能够很好地反映主题的发展状况。因此，微博中主题的时间序列表示如下：

$$Topic_k=[t_{k1},t_{k2},\cdots,t_{kN}] \tag{1}$$

式中，t_{ki} 表示主题 k 在第 i 个时间段内在微博帖子中出现的频率，N 表示时间序列长度。

2.2 微博中主题时序相似性计算

微博中，主题用户关注度的相似时序具有如下特征：1）相似时序具有形状相似的特点，其形状与幅度大小无关；2）相似性不受时序起始点所在位置的影响；3）在相等长度的时间内，时序的形状相似[2]。归一化的欧氏距离可识别同步的、形状相似的时序，滑动时间窗可动态获取定长的时序片段。因此，本文选用归一化的滑动时间窗欧氏距离[7]作为计算主题时序相似性的基本框架。当用户输入查询主题 Q 并选定感兴趣时间段（时间段长度为 m）后，相应地定义一个长度为 m 的滑动时间窗，计算每个窗口内的主题 R_k（$R_k\in$ 数据集$\{R_1,R_2,\cdots,R_M\}$）的子序列 $R_{k,l}=[t_{R_k(l+1)},t_{R_k(l+2)},\cdots,t_{R_k(l+m)}]$（$l$ 是窗口在 R_k 时序中的起始点）与查询序列 $Q=[t_{Q1},t_{Q2},\cdots,t_{Qm}]$ 的距离，即：

$$D(Q,R_{k,l})=\sqrt{\sum_{h=1}^{m}(\overline{t_{Qh}}-\overline{t_{R_k(l+h)}})^2} \tag{2}$$

式中，$\bar{t}$ 表示归一化后的时序频率值。

根据与查询主题相似值的大小对主题进行排序，从而为用户提供与其感兴趣主题时序相似的主题列表。

2.3 基于相关反馈技术的时序相似性度量的建立

上述检索结果能够帮助用户初步了解与其感兴趣主题相关的内容。引入相关反馈后，用户可根据自身兴趣进一步对每个结果给予是否相关的判断。查询方法可根据相关反馈结果进行学习以捕捉用户对特定时间点的喜好，进而学习出基于用户兴趣的相似性度量，为用户提供更满意的查询结果。本文通过反馈学习欧氏距离度量的权重，以获得集成用户深层次兴趣的距离度量。权重向量$\vec{w}$元素值的大小代表了用户对相应时间点的关注度在时序形状中作用的关注程度大小，此时主题时序相似性度量的计算公式如下：

$$D(Q,R_{k,l},\vec{w})=\sqrt{\sum_{i=1}^{m}w_i^2(\overline{t_{Qi}}-\overline{t_{R_k(l+i)}})^2} \tag{3}$$

式中，$\sum_{i=1}^{m}w_i^2=1$。

用户的相关反馈信息中包括用户感兴趣的时序和不感兴趣的时序两个方面，本文提出的查询方法通过分析用户感兴趣时序的特点识别出用户对查询序列中哪些点关注程度更高，从而发现用户对这些点更感兴趣；通过分析用户不感兴趣时序的特点，可以识别出用户对查询序列中哪些点关注程度较低，表明用户对这些点的感兴趣程度也相应较低。可见，正反馈时序和负反馈时序从不同角度描述了用户对相似主题时序中不同时间点的关注程度，反映了用户深层次的兴趣。基于以上分析，本文设计了一种综合考虑用户正、负反馈信息的优化目标函数，以建立体现用户兴趣的时序相似性度量。该目标函数旨在通过约束查询主题时序与用户判定相似的主题时序距离尽可能小，同时与用户判定不相似的主题时序距离尽可能大，来获取权重$\vec{w}$的最优解。目标函数表述如下：

$$L=\min_{\vec{w}}\left(\sum_{j=1}^{X}D(Q,R_{a(j)},\vec{w})-\sum_{v=1}^{Y}D(Q,R_{b(v)},\vec{w})\right) \tag{4}$$

式中，$X(0\leqslant X)$ 为用户标识为相似的检索结果数目；$Y(0\leqslant Y)$ 为用户判定为不相似的检索结果数目；$X+Y=H$，H 为用户选择观察的相似主题时序的数目；$a(j)$ 和 $b(v)$ 分别表示用户认为相似检索结果和不相似检索结果对应的数据集主题时序序号。

此优化问题，通过对目标函数中的变量 w_i 求偏导，可得到目标函数局部最优化的解，进而多次迭代后得到接近目标函数的全局最优解。

$$\frac{\partial L}{\partial w_i}=\sum_{j=1}^{X}\frac{2\times w_i\times(\overline{t_{Qi}}-\overline{t_{Ra(j)(l+i)}})^2}{\sqrt{\sum_{i=1}^{m}w_i^2(\overline{t_{Qi}}-\overline{t_{Ra(j)(l+i)}})^2}}-\sum_{v=1}^{Y}\frac{2\times w_i\times(\overline{t_{Qi}}-\overline{t_{Rb(v)(l+i)}})^2}{\sqrt{\sum_{i=1}^{m}w_i^2(\overline{t_{Qi}}-\overline{t_{Rb(v)(l+i)}})^2}} \tag{5}$$

综上所述，本文提出了一种基于相关反馈的微博相似主题时序查询方法，详细流程如算法 1 所示。

算法 1 基于相关反馈的微博相似主题时序查询

输入：查询主题 Q 及 m 个时间段，主题时间序列数据集$\{R_1,\cdots,R_M\}$，查询相似主题时序的个数 H。

输出：H 个与主题 Q 时序相似的主题时序。

算法过程如下：

步骤 1 统计 m 个时间段内查询主题 Q 对应的用户关注度时间序列。

for k=1 to M

步骤 2 利用尺寸为 m 的滑动时间窗和式(1)，计算查询主题 Q 的时序与数据集中主题 R_k 时序的距离。

end for

步骤 3 根据查询主题 Q 的时序与数据集中主题 R_k 时序的距离，从小到大排序数据集中的主题，并将前 H 个相似主题时序展示给用户。

步骤 4 根据用户反馈结果，利用梯度法迭代学习式(4)中的权重$\vec{w}$，迭代更新公式如下：

$$\vec{w}^{n+1}=\vec{w}^{n}-\beta\times\frac{\partial L}{\partial \vec{w}}$$

其中，$\vec{w}^{n}$ 表示第 n 次迭代的权重，β 为更新系数。

步骤 5　利用学习得到的权重和式(3)重复步骤 2、步骤 3。

步骤 6　若用户想继续检索，则继续步骤 4、步骤 5，否则停止。

3　实验结果和分析

3.1　数据描述

Twitter[1] 是微博客服务的代表性网站，平均每天有超过六千万的用户访问量和过亿的帖子更新[8]。我们通过 Twitter trends API[2] 对 2011 年 7 月 17 日至 2011 年 11 月 8 日期间 Twitter 中每天每小时的热点主题统计结果进行下载，共收集了 12480 个热点主题。在 Twitter 中，当主题的用户关注度达到一定程度时，主题将出现在热点主题列表中，Twitter 便将主题相关的帖子呈现给所有用户[9]。Twitter 中的热点主题每小时进行实时更新，可见用户的关注度与主题成为热点主题的频率相一致。因此，通过统计一段时间内主题出现在热点主题列表中的频率时间序列，可以展现此主题对应时间内的整体发展情况。本文以“天”为单位，并选择出现在热点主题列表中的频率总和不少于 28 的主题形成数据集，数据集中包含了 141 个热点主题。

3.2　微博相似时序主题查询系统

基于第 2 节中提出的方法，本文设计了一个微博相似主题时序查询的原型系统。该系统利用 Lucene[3] 为微博数据建立索引，用户可以输入需要查询的主题和相应的时间信息，系统根据用户输入信息查找该段时间内用户查询主题的用户关注度，作为系统当前的输入数据。利用 ChartDirector[4] 实现在系统中的时序变化曲线的显示。

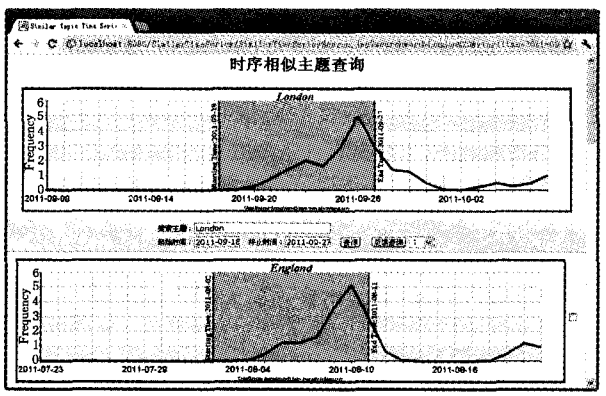

图 2　用户查询在“2011-09-18”至“2011-09-27”期间与主题“London”的相似主题时序查询系统界面

图 2 展示了微博相似主题时序查询系统的界面。用户选择“查询”与主题“London”在“2011-09-18”至“2011-09-27”间时序相似的“8”个主题。相似主题时序的时间段用绿色标识，显示对应时间段的起始点和终结点；每个检索结果的右侧有一个选项框，用户可根据自身需求对检索的结果给出相似与否的判断；当用户对检索结果给出反馈后，用户可通过“反馈查询”得到基于相关反馈实现体现用户兴趣的度量计算后的检索结果。

图 3 展示了基于归一化的滑动时间窗欧氏距离计算主题时序相似度的检索结果，用户对此检索结果排序中第 1 个、第 2 个、第 5 个结果表示满意，对其他结果表示不满意。从用户反馈信息可见，正反馈时序在起始点和终结点与查询时序有较高的相似性，且与查询时序第 7 天之后的形状比较相似；而负反馈时序在起始点与查询时序有较大的差异。另外，正、负反馈时序在第 5 天和第 6 天变化比较相似，可见用户对这两点的关注程度较低。根据用户正、负反馈时序的特点进行学习，获得了欧氏距离的权重向量 $w=[0.57\quad 0.32\quad 0.33\quad 0.24\quad 0.15\quad 0.16\quad 0.34\quad 0.20\quad 0.22\quad 0.39]$。向量 w 元素的最大值 0.57 为查询时序起始点的权重，第二大元素值 0.39 出现在查询时序终结点对应的天，查询时序第 7 天对应权重值为 0.34，第 5 天和第 6 天的权重值较小，分别为 0.15、0.16，与上述用户反馈结果的分析具有一致性。

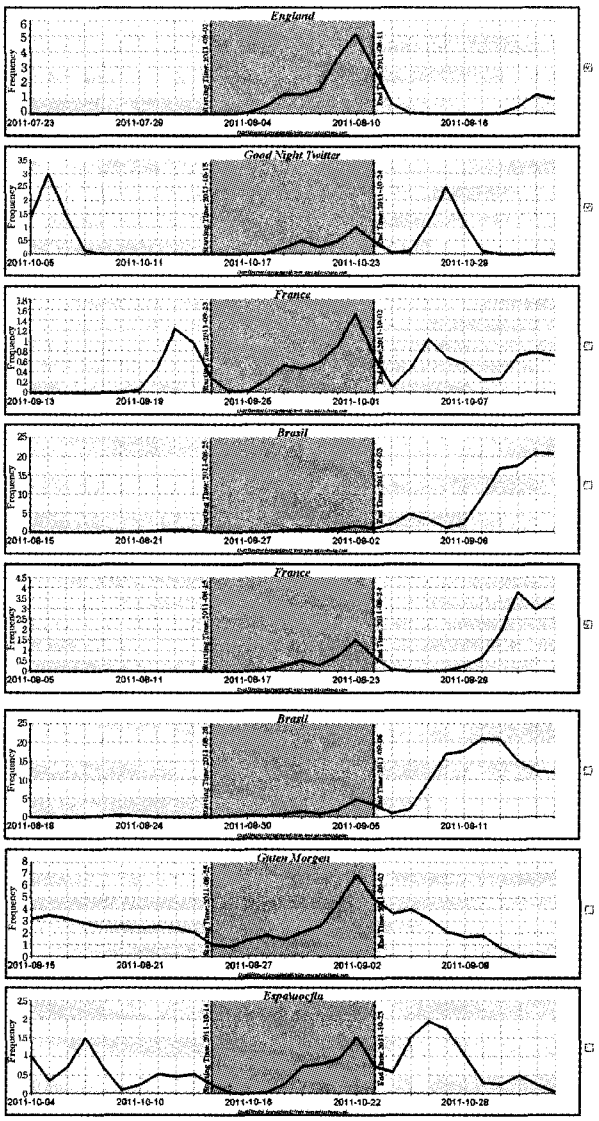

图 3　与主题“London”相似的主题时序查询结果展示(用户感兴趣时序以对号标示)

(下转第 198 页)

1) http://www.twitter.com/

2) https://dev.twitter.com/docs/api/1/get/trends/daily

3) http://lucene.apache.org/. Apache 软件基金会 4 jakarta 项目组的一个子项目，是一个开放源代码的全文检索引擎工具包

4) http://www.advsofteng.com/. 统计绘图工具

[J]. 自动化学报,2011,37(3):117-125
[12] 谢志强,邵侠,杨静. 存在设备无关延迟约束的综合柔性调度算法[J]. 机械工程学报,2011,47(4):181-189
[13] 刘世平,张洁,饶运清,等. 分布式车间管理控制系统研究[J]. 中国机械工程,2001,12(12):1432-435
[14] 包振强,李长仪,周鑫. 分布式混合优化调度方法研究[J]. 中国机械工程,2006,17(18):1908-1912
[15] Chung S H,Chan Felix T S,Chan H K. A modified genetic algorithm approach for scheduling of perfect maintenance in distributed production scheduling[J]. Engineering Applications of Artificial Intelligence,2009,22(7):1005-1014
[16] Oike S,Tanaka T. Robust production scheduling using autonomous distributed systems[J]. Key Engineering Materials,2012,516:166-169
[17] 谢志强,刘胜辉,乔佩利. 基于 ACPM 和 BFSM 的动态 Job-Shop 调度算法[J]. 计算机研究与发展,2003,40(7):79-85

(上接第 171 页)

图 4 展示了基于反馈的相似时序查询结果,用户标识第 1 个、第 2 个、第 4 个、第 5 个、第 8 个结果相似,可见,通过融合正、负反馈信息后的查询结果,提高了用户的满意度。另外,反馈后的检索结果中即使被用户标识为不相似的时序,也与初始查询结果中被用户标识为相似的时序匹配程度趋近一致,同时还与初始阶段被标识为不相似的主题时序在起始点有了明显区别。

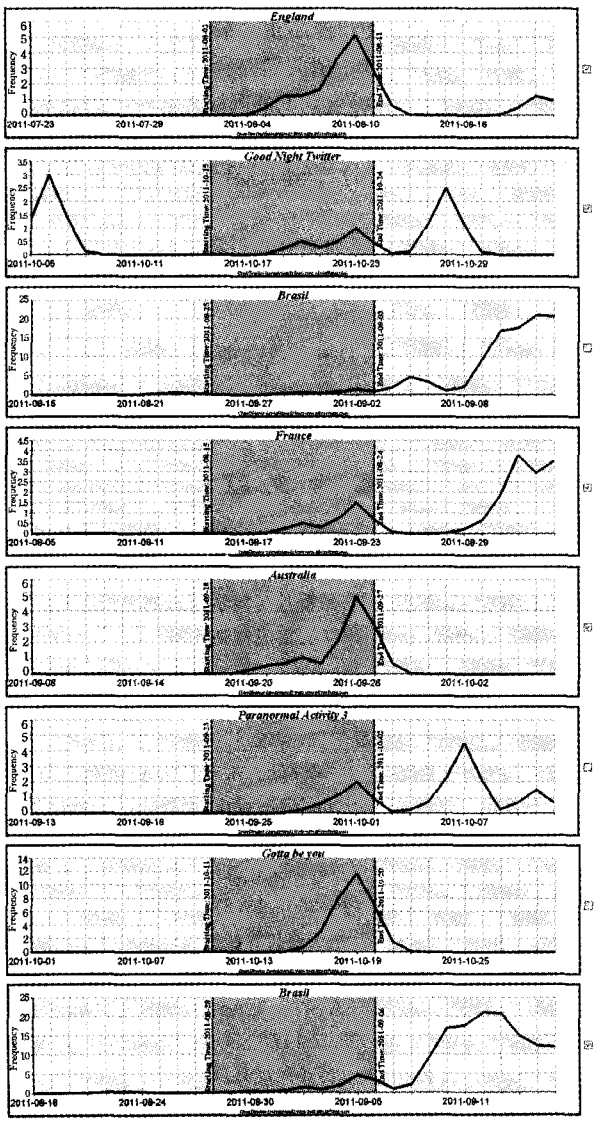

图 4 基于用户反馈的主题"London"相似时序结果展示(用户感兴趣时序以对号标示)

用户综合分析图 4 所示的时序相似主题"England"、"France"等在时序相似时间段后的发展变化情况,发现其具有"先快速下降继而平缓发展再上升"的发展趋势。基于该分析结果,用户进而可对查询主题"London"在"2011-09-27"之后的发展趋势进行预测,便于用户对其感兴趣的趋势(例如平缓发展)进行监控。

结束语 微博已经成为最受用户欢迎的了解网络热点主题发展状况的社会媒体平台之一,如何为微博用户提供相似主题时序查询也相应成为目前的研究热点。为了更好地帮助用户获取满足自身喜好的相似主题时序,本文提出了一种综合考虑正、负反馈信息的微博相似主题时序查询方法。实验结果表明,此方法能从用户标识为与查询时序是否相似的时序中挖掘出用户的兴趣,从而实现了体现用户偏好的时序相似性的计算,更好地为用户提供了满足其偏好的相似主题时序查询结果。未来将进一步考虑主题的内容信息,结合主题用户关注度时序为用户提供更全面的信息服务。

参 考 文 献

[1] He Y,Su W,Tian Y,et al. Summarizing microblogs on network hot topics[C]//iTAP:the 2011 International Conference on Internet Technology and Applications. 2011:1-4
[2] Yang J, Leskovec J. Patterns of temporal variation in online media[C]//WSDM'11:Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 2011:177-186
[3] Song S,Li Q,Bao H. Detecting dynamic association among twitter topics[C]//WWW 2012:Proceedings of the 2012 ACM Conference on the World Wide Web. 2012:605-606
[4] Keogh E J,Pazzani M J. Relevance feedback retrieval of time series data[C]//SIGIR 1999:the 22nd Annual ACM Conference on Special Interest Group on Information Retrieval. 1999:183-190
[5] 郑斌祥,席裕庚,杜秀华. 利用反馈的时序模式挖掘算法研究[J]. 控制与决策,2002,17(5):527-531
[6] 秦吉胜,王淑静,宋瀚涛. 基于小波变换和反馈的时间序列相似模式搜索算法[J]. 北京理工大学学报,2004,24(12):1069-1073
[7] Pawling A, Madey G. Feature Clustering for Data Steering in Dynamic Data Driven Application Systems[C]//ICCS 2009,Part II,Lecture Notes in Computer Science. Volume 5545,2009:460-469
[8] Meij E, Weerkamp W, Rijke M D. Adding Semantics to Microblog Posts[C]//WSDM'12:Proceedings of the fourth ACM international conference on Web search and data mining. 2012:563-572
[9] Griery C, Thomas K, Paxsony V, et al. @spam: The Underground on 140 Characters or Less[C]//CCS'10:Proceedings of the 17th ACM Conference on Computer and Communications Security. 2010:27-37