

局部多层网格划分下的轨迹数据概化方法

杨光*, 张磊, 李帆

(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

(*通信作者电子邮箱 yg397@126.com)

摘要:针对轨迹数据概化中空间划分的区域范围不能有效控制以及覆盖网格尺度难以合理选择的问题,提出局部多层网格划分方法,对样本密集的区域进行迭代划分。在此基础上提出一种轨迹数据概化方法,在局部多层网格划分的基础上,考虑时间约束合并轨迹连续往复通过的邻接区域,生成概化轨迹。真实数据的实验表明该算法得到的概化轨迹较同类算法保持了更多轨迹特性,更加适合后续数据挖掘,如聚类处理。

关键词:轨迹概化;局部多层网格;时间约束;轨迹数据

中图分类号:TP311 **文献标志码:**A

Trajectory data generalization based on local multi-hierarchy grid

YANG Guang*, ZHANG Lei, LI Fan

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou Jiangsu 221116, China)

Abstract: For current trajectory data generalization methods, the scope of the generalized regions cannot be controlled effectively, and the parameters of the grids can hardly be selected logically. This paper proposed the method of Local Multi-hierarchy Grid (LMG), so that the region with dense trajectory points would be divided iteratively. And then a method for trajectory data generalization named TRAGenLMG was proposed, which was based on LMG, and time-constraint was used to merge some adjoining grids, and finally the generalized trajectory was got. The experiments with real open dataset show that the generalized trajectories generated by TRAGenLMG can well maintain the temporal feature of the trajectory data and can be efficiently applied into further data analysis.

Key words: trajectory generalization; Local Multi-hierarchy Grid (LMG); time-constraint; trajectory data

0 引言

移动对象空间定位技术的发展产生丰富的轨迹数据,海量轨迹数据处理过程耗时久,占用空间大。对轨迹数据进行概化,能够在保持数据原有位置等特征信息的基础上对轨迹进行抽象化处理,缩减数据的规模,有助于海量轨迹数据的高效处理。轨迹概化一般建立在对轨迹空间进行区域划分的基础上。文献[1]使用基于密度的方法对轨迹点分组,根据分组结果构建 Voronoi 图进行区域划分。在我们的前期工作^[2]中,首先根据时间、空间位置、速度等因素提取轨迹的特征点,然后对特征点进行聚类,得到的聚类簇作为概化的区域分组。文献[3]在进行轨迹数据热点区域发现时用网格划分轨迹数据的样本空间,网格由一系列规则的小正方形单元格组成。现有移动对象运动空间划分方法中,对于划分样本空间网格单元格的尺度选择没有一个明确的依据,通常需要事先了解样本的采样率、样本数据的运动范围、移动对象运动的速度等因素,综合考虑这些因素主观地选择一个划分尺度。这种盲目行为对挖掘结果的准确度和可行度都有很大影响。

针对轨迹数据概化中区域划分的区域范围不能有效控制以及覆盖网格尺度不能合理选择的问题,提出局部多层次划分方法,对于高密度样本区域迭代地进行多层次网格划分。在网格划分的基础上,提出了局部多层网格下的轨迹数据概化方法。检查轨迹经过的区域网格,在时间约束下进行局域

合并,生成概化轨迹。

1 局部多层网格划分方法

局部多层网格划分(Local Multi-hierarchy Grid, LMG)方法对于高密度样本区域迭代地进行多层次划分,每深入一个层次的划分都使覆盖网格的精度提高一倍,最终使得覆盖高密度区域的网格单元格尺寸能取到一个合适值。因此算法中对于覆盖网格单元格尺寸的初始值不会对结果产生重大影响。而且通常情况下,为了提高算法的效率,单元格的初始尺寸都会选择一个比较大的值,例如初始时仅仅对样本空间四分,然后在此基础上做进一步的迭代划分。对样本空间采用灵活的划分尺度并利用这些细分单元格进行轨迹概化,有效避免得到区域范围过大的问题,也在一定程度上减弱了网格划分对样本数据点之间联系产生的割裂影响。LMG 是一个递归过程,如算法 1。

算法 1 LMG(D, d, n)

输入 轨迹样本点集合 D ; 划分初始参数 d ; 每个单元格内最大数据点个数 n 。

输出 局部多层次划分的网格 G 。

1) 把样本空间划分为 $d \times d$ 个相同的矩形单元格 $\{c_{0i} | 0 \leq i \leq (d \times d)\}$, 将这些初始划分得到的单元格加入需要进一步划分的单元格 C 中。

2) 统计每个单元格中点的个数 $count(c_{0i})$, 如果单元格

收稿日期:2012-12-27;修回日期:2013-02-18。 基金项目:教育部博士点基金资助项目(20110095110010);江苏省博士后基金资助项目(0802023C);中国矿业大学青年科技基金资助项目(2008A040);中国矿业大学研究生实践与科研创新课题专项基金资助项目(GSF122111)。

作者简介:杨光(1988-),女,山东济宁人,硕士研究生,主要研究方向:移动对象轨迹数据挖掘;张磊(1977-),男,江苏沛县人,副教授,博士,主要研究方向:移动对象轨迹数据挖掘;李帆(1986-),男,四川南充人,硕士,主要研究方向:移动对象轨迹数据挖掘。

$c_{0,i}$ 中数据点个数大于 n , 则将该单元格加入集合 C_0 中, 否则将该单元格加入集合 G 中。

3) 对于 C_0 中的每个单元格 $c_{0,i}$, 重复步骤 2)。

2 局部多层网格划分下的轨迹概化方法

局部多层网格划分下的轨迹概化 (Trajectory Data Generalization Based on Local Multi-hierarchy Grid, TRAGenLMG) 方法, 在轨迹概化过程中应用 LMG, 样本密集的区域被多次划分, 使得划分的单元格非常小, 覆盖单元样本点稀疏区域的网格则较大, 得到的原始轨迹点各个分组所包含的轨迹点的个数基本一致。在网格划分的基础上, 检查轨迹经过的区域, 在时间约束下进行局域合并, 最终生成概化轨迹。

2.1 时间约束下的概化区域合并

现有轨迹数据概化方法主要考虑的是空间位置的聚集, 在时间维度上则考虑得较少。移动对象的活动是“时空”维度上的整体^[5], 因此在进行轨迹概化的时候, 轨迹的时间属性是必须要考虑的因素之一。

考虑时间属性的轨迹概化存在轨迹交叉重叠的问题。移动对象可能会在一段时间段内在若干个邻接区域之间做往复运动, 如图 1, 联系紧密的轨迹点可能会被划分的网格割裂成不同区域。原始的区域划分方法生成的概化轨迹多次连续通过若干个邻接区域, 造成结果数据冗余、概化结果不精确。对于 LMG 生成的网格, 考虑时间约束进行区域合并。在生成初始概化轨迹后, 对初始概化轨迹进行路径检查, 统计概化轨迹通过的区域。如果发现概化轨迹在连续时间内多次通过若干个邻接区域, 那么将这些邻接区域合并, 重新计算代表点的位置坐标, 生成概化轨迹。

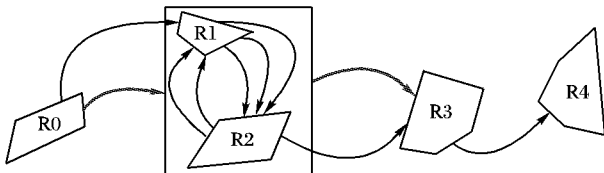


图1 被合并的概化区域

在进行初始概化轨迹的路径检查时候需要判断两个单元格是否相邻, 如图 2。对于单元格 c_1, c_2 , 假如 c_1 的中心点为 p_1 , 宽为 w_1 , 高为 h_1 ; c_2 中心点为 p_2 , 宽为 w_2 , 高为 h_2 。图中 (b), (c) 两种情况可以看作是 (a) 的特殊情况, 它们都是 (a) 中矩形在一边上运动产生的重合; 而且, 如果两个矩形的宽或高相等, 那么 (b), (c) 两种情况继续运动后产生情况 (d), (e)。可以从横向和纵向两个维度来判断两个矩形是否邻接, 如果 $(w_1 + w_2)/2 < |p_1.x - p_2.x|$, 则矩形横向分离, 如果 $(h_1 + h_2)/2 < |p_1.y - p_2.y|$, 则矩形纵向分离, 如果两个矩形既不横向分离又不纵向分离, 则可认定两个矩形相邻或相接。

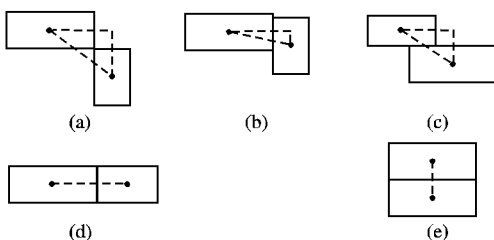


图2 单元格相邻接的判断

2.2 TRAGenLMG 的整体步骤

算法 TRAGenLMG 描述如下:

算法 2 TRAGenLMG(D, d, n)。

输入 轨迹数据集 D , 初始划分参数 d , 每个单元格内最大数据点个数 n ;

输出 概化轨迹数据集 DG 。

- 1) 从轨迹数据集 D 中提取轨迹点集合 P ;
- 2) 局部多层次网格划分 LMG(P, d, n), 得到网格集合 S ;
- 3) 对于 DG 中的每一条轨迹:
- 4) 统计原始轨迹路经过的网格信息, 查找原始轨迹经过的一系列网格 S' ;
- 5) 检查 S' , 如果 S' 中存在若干个邻接区域, 轨迹在连续时间内多次通过这些区域, 那么将这些邻接区域从集合 S' 中删除;
- 6) 将这些邻接区域合并, 重新计算合并后网格区域中心;
- 7) 将合并后的区域加入集合 S' 中;
- 8) 用网格区域中心的连接来代替原始轨迹, 生成概化轨迹, 将其放入概化轨迹数据集 DG 。

3 实验及分析

针对轨迹数据的研究, 开发了轨迹数据智能分析平台 TrajectoryMinerSystem, 用于验证本文方法的是轨迹数据概化模块 TRAGEN, 选取真实公测数据集飓风数据^[7]进行实验。轨迹概化是对轨迹数据抽象化处理的过程, 会带来信息丢失。在概化过程中将信息丢失控制在一定范围内, 从而保证原始轨迹信息的有效性。文献[8]根据空间位置信息衡量轨迹信息丢失 (Information Loss, IL)。接着, 以轨迹聚类为例证明概化轨迹在后续轨迹处理中的可用性和有效性。在 TRAGenLMG 的基础上, 将空间概化所得到的轨迹数据作为数据源进行聚类^[9]。在控制参数相同的情况下, 比较概化轨迹聚类和原始轨迹聚类的效果。根据集合相似性计算函数, 采用不同聚类结果的定量比较方法^[2], 得到聚类结果的相似性 (ClusterSimilarity)。在以下实验中用于验证概化效果的聚类参数是一致的。

3.1 LMG 和均匀网格划分对概化结果的对比

TRAGenLMG 方法中使用 LMG 对高密度区域进行迭代划分, 不同初始划分参数下, 概化轨迹的形状大致相同, 且轨迹点分布密集的地方概化轨迹对于原始轨迹的拟合程度更佳, 这样生成的概化轨迹更好地保留了较多数轨迹点的原始特性。而对于轨迹点分布稀疏的地区, 概化轨迹进行一定的取舍和降维。使用均匀网格划分得到的概化结果, 对于同样的样本, 当单元格划分比较细密的时候, 概化轨迹能得到较好的效果。随着网格初始划分参数减小, 网格尺度增大, 划分越来越粗糙, 这或许能降低时间复杂度, 但是结果的准确性却严重下降。

综上, 使用 LMG 划分网格覆盖运动平面的自适应效果比较好, 只要设定单元格内样本点个数阈值, 初始划分参数对概化结果不会有太大影响; 使用均匀网格划分, 结果对单元格尺度变化比较敏感。

3.2 LMG 和均匀网格划分效率的对比

比较 LMG 与均匀网格划分方法的效率和结果, 使用飓风数据 (2001—2009 年, 共 113 条轨迹, 4 178 个轨迹点), 在 LMG 中迭代划分的条件是单元格中的样本点超过 20。

图3(a)是LMG在不同初始划分参数下,得到划分网格单元格的数量和相应的划分时间。虽然初始划分参数不断增大,但是划分所得单元格数量和网格划分时间并没有明显增加。从划分过程来看,使用网格划分运动空间并统计每个单元格的样本点数量以后算法的时间复杂度与网格单元格的数目有关,因此,较稳定的划分方式是保证算法时间消耗保持稳定的重要条件之一。同时,使用LMG划分运动平面,虽然初始划分参数不同,但是网格数量保持较稳定,从而保证在同等轨迹点数目阈值下轨迹概化结果保持稳定。

作为对比实验,图3(b)显示,使用单层次均匀划分方法,在相同的空间范围,单元格尺寸越小,划分后得到的网格越多,网格划分消耗的时间也越多;单元格尺寸越大,概化轨迹误差越大,且概化轨迹的有效特征减少。因此该方法对参数的适应性较差。

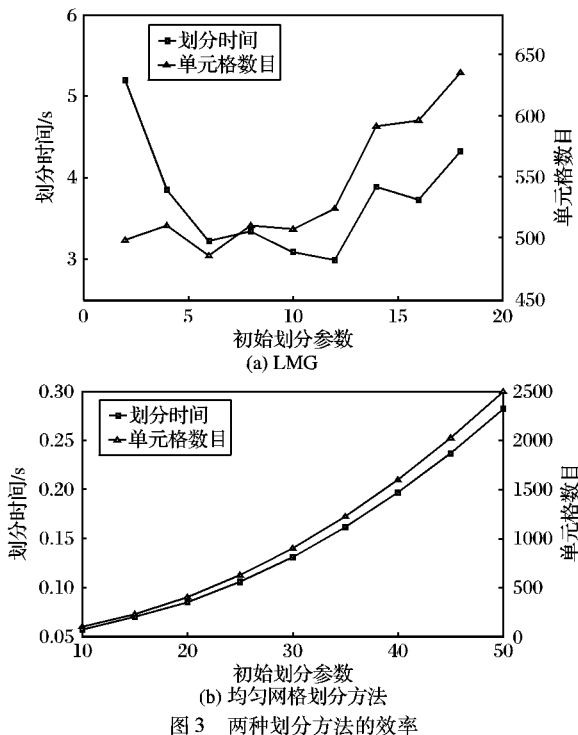


图3 两种划分方法的效率

3.3 样本规模及参数设置对于算法的影响

考察参数设置对 TRAGenLMG 算法结果的影响,实验使用飓风数据(2001—2009年,共113条轨迹,4178个轨迹点)。图4(a)中,LMG迭代划分的条件是单元格中的样本点超过20;图4(b)中,初始化法参数设置为2。从信息丢失 IL 和轨迹聚类 $ClusterSimilarity$ 两个方面衡量概化效果。

考察初始划分参数对于概化效果的影响。图4(a)中,随着初始划分参数不断增大,概化轨迹的 IL 和 $ClusterSimilarity$ 并没有明显增加或减少,而是在一定区间内波动。3.2节中已经证明初始化法参数对于网格划分的结果以及生成网格数目影响不大,因此对概化轨迹的结果影响也较小。接着,考察单元格内轨迹点个数阈值对于概化结果的影响,如图4(b)所示,随着轨迹点个数阈值的减小, IL 逐渐减小, $ClusterSimilarity$ 逐渐增大。这是因为轨迹点个数阈值设置越小,网格划分就越细致,从而得到的概化轨迹对于原始轨迹的拟合程度就越高。这样的概化轨迹保持了更多原始轨迹的信息,应用在后续挖掘中得到的结果与原始轨迹相似度也越高。

考察样本规模对于概化效果的影响,使用飓风数据

(1980—2009年,以3年为单位逐次增加数据)。在LMG中,初始划分参数设置为2,LMG中轨迹点个数阈值为20。

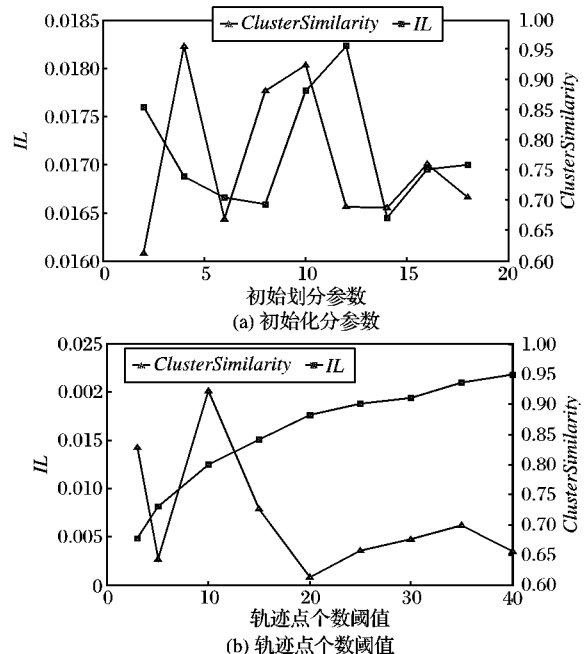


图4 参数设置对概化效果的影响

如图5所示,随着轨迹点个数的增加,生成的概化轨迹与原始轨迹比较, IL 逐渐减小,而 $ClusterSimilarity$ 逐渐增加。当轨迹点数目大于4000, $ClusterSimilarity$ 保持在0.8以上。这说明随着样本规模的增加,在初始化法参数不变的情况下,概化轨迹的各种特性更加接近原始轨迹,也更加适应后续挖掘处理。

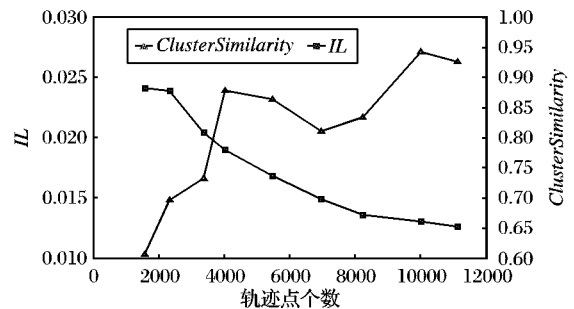


图5 样本规模对概化结果的影响

图6比较概化轨迹和原始轨迹在聚类处理中时间消耗情况,随着样本规模的增加,原始轨迹和概化轨迹的聚类时间都不断增大,然而概化轨迹的聚类时间始终小于原始轨迹的聚类时间。

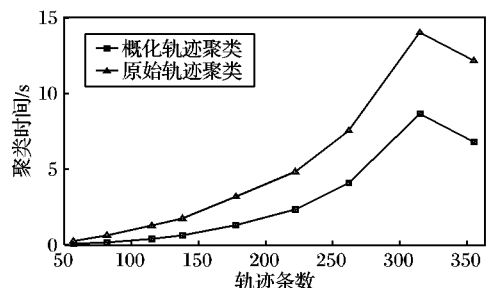


图6 概化对轨迹聚类时间的影响

综上所述,TRAGenLMG得到的概化轨迹,一方面很好地保持了原始轨迹的时空特性,另一方面可以有效应用于轨迹聚类等后续挖掘中,具有较高的时间效率。

4 结语

本文研究轨迹数据的概化方法,针对轨迹数据概化中区域划分的区域范围不能有效控制以及覆盖网格尺度不能合理选择的问题,提出了 LMG 方法来多次划分样本密集的区域,使各个区域之间的轨迹点密度基本一致。在此基础上提出了 TRAGenLMG 方法,在时间约束下合并连续往复通过的邻接区域,生成概化轨迹。使用真实数据集进行实验,从信息丢失量和概化轨迹聚类与原始轨迹聚类的结果相似性两个方面进行概化效果的衡量。实验结果显示:TRAGenLMG 在一定程度上保持了原始轨迹信息的同时,对于后续挖掘处理具有较好的适用性,效率相对于原始轨迹较高。

参考文献:

- [1] ANDRIENKO N, ANDRIENKO G. Spatial generalization and aggregation of massive movement data[J]. *Visualization and Computer Graphics*, 2011, 17(2): 205–219.
- [2] ZHANG L, YANG G, WANG Z C. Trajectory clustering based on spatial generalization[J]. *Journal of Information & Computational Science*, 2012, 9(2): 315–32.
- [3] GIANNOTTI F, NANNI M, PEDRESCHI D, *et al.* Trajectory pattern mining[C]// *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2007: 330–339.
- [4] 唐良, 唐常杰, 姜页希, 等. TRAODGrid: 基于 Grid 空间划分的高效离群轨迹检测方法[J]. *计算机研究与发展*, 2008, 45(10): 185–190.
- [5] SHERKAT R, LI J, MAMOULIS N. Efficient time stamped event sequence anonymization[EB/OL]. [2012-08-20]. <http://www.cs.hku.hk/research/techreps/document/TR-2011-02.pdf>.
- [6] ASSAM R, SEIDL T. Preserving privacy of moving objects via temporal clustering of spatio-temporal data streams[C]// *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. New York: ACM, 2011: 9–16.
- [7] Unisys Weather [DB/OL]. [2012-05-23]. <http://weather.unisys.com/hurricane/atlantic/>.
- [8] STEFANAKIS E. Trajectory generalization under space constraints[EB/OL]. [2012-08-20]. http://www.giscience.org/proceedings/abstracts/giscience2012_paper_74.pdf.
- [9] MASCIARI E. A framework for trajectory clustering[C]// *Proceedings of the 3rd International Conference on GeoSensor Networks*. Berlin: Springer-Verlag, 2009: 102–111.
- [10] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. *软件学报*, 2008, 19(1): 48–61.
- [11] 袁冠, 夏士雄, 张磊, 等. 基于结构相似度的轨迹聚类算法[J]. *通信学报*, 2011, 32(9): 103–110.
- [12] MARTINEZ-BEA S. Trajectory anonymization from a time series perspective[C]// *2011 IEEE International Conference on Fuzzy Systems*. Piscataway: IEEE, 2011: 401–408.
- [13] ANDRIENKO G, ANDRIENKO N, GIANNOTTI F, *et al.* Movement data anonymity through generalization[J]. *Journal of Transactions on Data Privacy*, 2010, 3(2): 91–121.
- [39] HATZIVASSILOGLU V, WIEBE J. Effects of adjective orientation and gradability on sentence subjectivity [C]// *Proceedings of International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2000: 299–305.
- [40] HATZIVASSILOGLU V, KLAIVANS J L, HOLCOMBE M L, *et al.* SIMFINDER: A flexible clustering tool for summarization [C]// *Proceedings of the Workshop on Summarization in NAACL-01*. Stroudsburg, PA: Association for Computational Linguistics, 2001: 41–49.
- [41] BENAMARA F, CHARDON B, MATHIEU Y, *et al.* Towards context-based subjectivity analysis [C]// *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011: 1180–1188.
- [42] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]// *Proceedings of Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2002: 417–424.
- [43] KIM S M, HOVY E. Determining the sentiment of opinions[C]// *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1367–1373.
- [44] PANG B, LEE L. Opinion mining and sentiment analysis[J]. *Journal Foundations and Trends in Information Retrieval*, 2008, 2(2): 1–135.
- [45] CHOI Y, CARDIE C, RILOFF E. Identifying sources of opinions with conditional random fields and extraction patterns [C]// *Proceedings of HLT/EMNLP-2005*. Stroudsburg, PA: Association for Computational Linguistics, 2005: 355–362.
- [46] BETHARD S, YU H, THORNTON A. Automatic extraction of opinion propositions and their holders [C]// *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*. Stanford: American Association for Artificial Intelligence, 2004: 22–24.
- [47] PANG B, LEE L. Using very simple statistics for review search: An exploration [C]// *Proceedings of International Conference on Computational Linguistics*. Manchester, UK: Coling 2008 Organizing Committee, 2008: 75–78.
- [48] DAVE K, LAWRENCE S, PENNOCK D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C]// *Proceedings of the 12th International World Wide Web Conference*. New York: ACM, 2003: 519–528.
- [49] LIU B, HU M, CHENG J. Opinion observer: analyzing and comparing opinions on the Web [C]// *Proceedings of the 14th International Conference on World Wide Web*. New York: ACM, 2005: 342–351.
- [50] WILSON T, HOFFMANN P, SOMASUNDARAN S, *et al.* OpinionFinder: A system for subjectivity analysis [C]// *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics, 2005: 34–35.
- [51] GAMON M, AUE A, CORSTON-OLIVER S, *et al.* Pulse: Mining customer opinions from free text [C]// *Proceedings of the 6th International Symposium on Intelligent Data Analysis*. Berlin: Springer-Verlag, 2005: 121–132.
- [52] CHERRY C, MOHAMMAD S. Binary classifiers and latent sequence models for emotion detection in suicide notes [J]. *Journal of Biomedical Informatics Insights*, 2012, 5(S1): 147–154.

(上接第 1578 页)