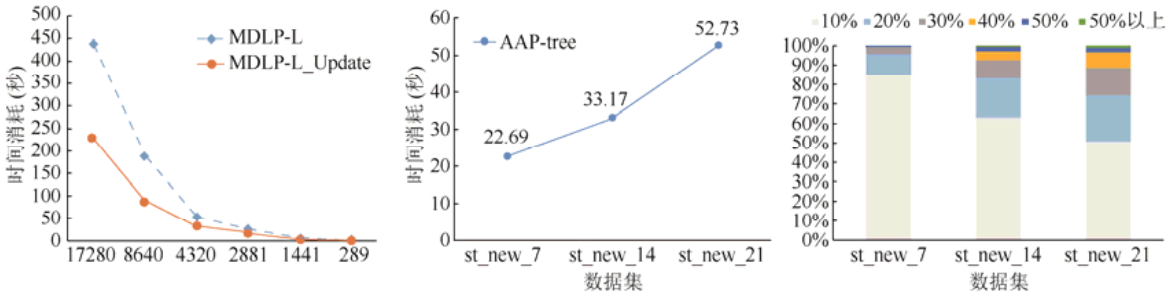


(a) 时间轴划分更新效率 (b) AAP-Tree 更新效率 (c) 更新停留点集的时刻覆盖数量

Fig.9 The efficiency of updating time axis dividing and AAP-tree and the distribution of new stops on time axis

图 9 时间轴、AAP-tree 更新效率与更新停留点集时刻覆盖

由图 8 和图 9 可以看出,本文方法无论在数据预处理、索引建立、频繁模式挖掘还是时间轴划分与索引的更新方面都是高效的,同时,本文方法相比现有方法还具有一定的灵活性.

6 总结及展望

挖掘语义轨迹的频繁模式是实现旅游线路推荐、路线预测、用户生活模式挖掘、朋友推荐等应用的技术基础,在很多情况下,用户会对这些应用有到达时间限制的需求,而现有的语义轨迹方法大多没有考虑到达时间限制,少数考虑了到达时间限制却因为限制太强而导致挖掘不到频繁的模式.因此,本文首次研究并形式化定义了语义轨迹的 AAFP,并使用信息熵聚类的方法将语义轨迹集中各个地点的时间轴合理地划分开,并提出了挖掘语义轨迹的 AAFP 的基线算法.之后,为了改进基线算法使其更高效、更灵活,本文建立了一个多层混合索引 AAP-tree,并给出了基于其上的语义轨迹 AAFP 挖掘算法.然后,针对新数据到来后的维护问题,提出了时间轴划分及 AAP-tree 的高效维护方案.最后,在真实数据集上进行实验,实验结果证明了本文方法的有效性与高效性.

在未来的工作中,我们还将根据实际中存在的新颖性需求对数据集进行维护,比如,商场的管理者会希望仅维护近 3 个月的数据集以确保数据能够反映最近的流行趋势,在这样的场景下,本文方法还需在维护时间轴划分与索引时考虑删除点的情况.另外,本文方法主要是针对挖掘频繁的模式,设定频繁阈值可以过滤掉大量的不符合需求的模式以加快挖掘速度,若需要挖掘所有的模式,则本文方法并不适用,故未来还需考虑高效的模式挖掘方法.

References:

[1] Zheng Y. Trajectory data mining: An overview. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(3):29. [doi: 10.1145/2743025]

[2] Ying JJC, Lee WC, Weng TC, Tseng VS. Semantic trajectory mining for location prediction. In: Agrawal D, Cruz I, Jensen CS, Ofek E, Tanin E, eds. Proc. of the 19th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2011. 34–43. [doi: 10.1145/2093973.2093980]

[3] Ying JC, Chen HS, Lin KW, Lu HC, Tseng VS, Tsai HW, Cheng KH, Lin SC. Semantic trajectory-based high utility item recommendation system. Expert Systems with Applications, 2014,41(10):4762–4776. [doi: 10.1016/j.eswa.2014.01.042]

[4] Li Q, Zheng Y, Xie X, Chen YK, Liu WY, Ma WY. Mining user similarity based on location history. In: Aref WG, Mokbel MF, Schneider M, eds. Proc. of the 16th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2008. 34. [doi: 10.1145/1463434.1463477]

[5] Zheng Y, Zhang L, Ma Z, Xie X, Ma WY. Recommending friends and locations based on individual location history. ACM Trans. on the Web (TWEB), 2011,5(1):5. [doi: 10.1145/1921591.1921596]

[6] Monreale A, Pinelli F, Trasarti R, Giannotti F. Wherenext: A location predictor on trajectory pattern mining. In: Elder J, Fogelman FS, eds. Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 637–646. [doi: 10.1145/1557019.1557091]

[7] Zhang C, Han J, Shou L, Lu J, Porta TL. Splitter: Mining fine-grained sequential patterns in semantic trajectories. Proc. of the VLDB Endowment, 2014,7(9):769–780. [doi: 10.14778/2732939.2732949]

[8] Chen CC, Kuo CH, Peng WC. Mining spatial-temporal semantic trajectory patterns from raw trajectories. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining Workshop (ICDMW). New York: IEEE, 2015. 1019–1024. [doi: 10.1109/ICDMW.2015.55]

[9] Jeung H, Yiu ML, Jensen CS. Trajectory pattern mining. In: Computing with Spatial Trajectories. Berlin, Heidelberg: Springer-Verlag, 2011. 143–177. [doi: 10.1007/978-1-4614-1629-6]

[10] Aggarwal CC, Han JW. Frequent Pattern Mining. New York: Springer-Verlag, 2014. [doi: 10.1007/978-3-319-07821-2]

[11] Baglioni M, Renso C, Trasarti R, Wachowicz M. Towards semantic interpretation of movement behavior. In: Sester M, Bernard L, Paelke V, eds. Advances in GIScience. Lecture Notes in Geoinformation and Cartography, Berlin, Heidelberg: Springer-Verlag, 2009. 271–288. [doi: 10.1007/978-3-642-00318-9_14]

[12] Alvares LO, Bogorny V, Kuijpers B, Macedo JAFD, Moelans B, Vaisman A. A model for enriching trajectories with semantic geographical information. In: Samet H, Shahabi C, eds. Proc. of the 15th Annual ACM Int'l Symp. on Advances in Geographic Information Systems. New York: ACM Press, 2007. 1–8. [doi: 10.1145/1341012.1341041]

[13] Giannotti F, Nanni M, Pinelli F, Pedreschi D. Trajectory pattern mining. In: Berkhin P, ed. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 330–339. [doi: 10.1145/1281192.1281230]

[14] Zheng K, Zheng Y, Yuan NJ, Shang S, Zhou X. On discovery of gathering patterns from trajectories. In: Proc. of the 29th IEEE Int'l Conf. on Data Engineering (ICDE). New York: IEEE, 2013. 242–253. [doi: 10.1109/ICDE.2013.6544829]

[15] Hung CC, Peng WC, Lee WC. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. The Int'l Journal on Very Large Data Bases, 2015,24(2):169–192. [doi: 10.1007/s00778-011-0262-6]

[16] Matsubara Y, Li L, Papalexakis E, Lo D, Sakurai Y, Faloutsos C. F-Trail: Finding patterns in taxi trajectories. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, eds. Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer-Verlag, 2013. 86–98. [doi: 10.1007/978-3-642-37453-1_8]

[17] Roh GP, Roh JW, Hwang SW, Yi BK. Supporting pattern-matching queries over trajectories on road networks. IEEE Trans. on Knowledge and Data Engineering, 2011,23(11):1753–1758. [doi: 10.1109/TKDE.2010.189]

[18] Roh GP, Hwang S. TPM: Supporting pattern matching queries for road-network trajectory data. In: Ailamaki A, Amer-Yahia S, Pate J, Risch T, Senellart P, Stoyanovich J, eds. Proc. of the 14th Int'l Conference on Extending Database Technology. New York: ACM Press, 2011. 554–557. [doi: 10.1145/1951365.1951439]

[19] Qiu M, Pi D. Mining frequent trajectory patterns in road network based on similar trajectory. In: Yin H, ed. Proc. of the Intelligent Data Engineering and Automated Learning (IDEAL 2016). Lecture Notes in Computer Science, 2016. 46–57. [doi: 10.1007/978-3-319-46257-8_6]

[20] Fayyad UM. Multi-Interval discretization of continuous-valued attributes for classification learning. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 1993. 1022–1027.

[21] Kim Y, Han J, Yuan C. TOPTRAC: Topical trajectory pattern mining. In: Cao LB, Zhang CQ, eds. Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2015. 587–596. [doi: 10.1145/2783258.2783342]

附录

A.1 信息熵增量公式推导

根据第 3.1 节,地点 l 停留点分布的信息熵计算公式为 $Ent(S_l)=-\sum_{i=1}^{k}p(C_i)\log(p(C_i))=-\sum_{i=1}^{k}\frac{|C_i|}{|S_l|}\log\left(\frac{|C_i|}{|S_l|}\right)$,其中,$1\leqslant i\leqslant k$.当一条新的数据 s_{new} 到来时,设 $s_{new}\in C_j$,$1\leqslant j\leqslant k$,信息熵的变化如下:

$$Ent(S_l)_{new}=-\left(\sum_{j-1}^{i=1}\left(\frac{|C_i|}{|S_l|+1}\log\left(\frac{|C_i|}{|S_l|+1}\right)\right)+\frac{|C_j|+1}{|S_l|+1}\log\left(\frac{|C_j|+1}{|S_l|+1}\right)+\sum_{i=j+1}^{k}\left(\frac{|C_i|}{|S_l|+1}\log\left(\frac{|C_i|}{|S_l|+1}\right)\right)\right).$$

根据比值的对数函数可以表示为其分子分母对数函数的差的性质,上述两个公式可以变形为

$$Ent(S_l)=-\sum_{i=1}^{k}\frac{|C_i|}{|S_l|}(\log(|C_i|)-\log(|S_l|)) \tag{5}$$

$$\left.\begin{aligned}Ent(S_l)_{new}=-\left(\sum_{j-1}^{i=1}\left(\frac{|C_i|}{|S_l|+1}(\log(|C_i|)-\log(|S_l|+1)\right)+\frac{|C_j|+1}{|S_l|+1}\left(\log(|C_j|+1)-\log(|S_l|+1)\right)+\right.\\ \left.\sum_{i=j+1}^{k}\left(\frac{|C_i|}{|S_l|+1}\left(\log(|C_i|)-\log(|S_l|+1)\right)\right)\right)\end{aligned}\right\} \tag{6}$$

用式(6)减去式(5)得到的分母与分子如下:

$$\left.\begin{aligned}&den\left(Ent(S_l)_{new}-Ent(S_l)\right)=-(|S_l|+1)|S_l|num\left(Ent(S_l)_{new}-Ent(S_l)\right)=\left[\sum_{j-1}^{i=1}|S_l||C_i|\log(|C_i|)+|S_l||C_j|\log(|C_j|+1)+\right.\\ &\left.\sum_{i=j+1}^{k}|S_l||C_i|\log(|C_i|)-\underline{\sum_{i=1}^{k}|S_l||C_i|\log(|S_l|+1)}+|S_l|\log(|C_j|+1)-|S_l|\log(|S_l|+1)\right]\\ &-\left[\sum_{i=1}^{j-1}|S_l||C_i|\log(|C_i|)+|S_l||C_j|\log(|C_j|)+\sum_{i=j+1}^{k}|S_l||C_i|\log(|C_i|)-\underline{\sum_{i=1}^{k}|S_l||C_i|\log(|S_l|)}-\sum_{i=1}^{k}(|C_i|\log(|C_i|)-|C_i|\log(|S_l|))\right]\end{aligned}\right\} \tag{7}$$

式(7)中,标记为灰色的部分可以抵消,标记为下划线的部分相加后可得:

$$-\sum_{i=1}^{k}|S_l||C_i|\log\left(\frac{|S_l|+1}{|S_l|}\right)=-|S_l|\cdot\log\left(\frac{|S_l|+1}{|S_l|}\right)\cdot\sum_{i=1}^{k}|C_i|=-|S_l|^2\cdot\log\left(\frac{|S_l|+1}{|S_l|}\right) \tag{8}$$

式(7)中余下部分整理后可得:

$$\left.\begin{aligned}&|S_l||C_j|\log(|C_j|+1)+|S_l|\log(|C_j|+1)-|S_l|\log(|S_l|+1)-|S_l||C_j|\log(|C_j|)-\sum_{i=1}^{k}(|C_i|\log(|C_i|)-|C_i|\log(|S_l|))\\ &=|S_l||C_j|\log\left(\frac{|C_j|+1}{|C_j|}\right)+|S_l|\log\left(\frac{|C_j|+1}{|S_l|+1}\right)-|S_l|\cdot\sum_{i=1}^{k}\frac{|C_i|}{|S_l|}\log\left(\frac{|C_i|}{|S_l|}\right)\\ &=|S_l||C_j|\log\left(\frac{|C_j|+1}{|C_j|}\right)+|S_l|\log\left(\frac{|C_j|+1}{|S_l|+1}\right)-|S_l|Ent(S_l)\end{aligned}\right\} \tag{9}$$

合并式(8)、式(9),并将分母分子合并即可得:

$$Ent(S_l)_{new}-Ent(S_l)=\frac{|S_l|\cdot\log\left(\frac{|S_l|+1}{|S_l|}\right)-|C_j|\cdot\log\left(\frac{|C_j|+1}{|C_j|}\right)-\log\left(\frac{|C_j|+1}{|S_l|+1}\right)+Ent(S_l)}{|S_l|+1}.$$

A1.2 信息熵批量增量公式推导

对于地点 l,假设新到来 m 个点,这些点非平均地分布在 s 个类中:$\bar{C}_1,\bar{C}_2,\ldots,\bar{C}_s$,每个类对应的新增点数量为 $a_1,a_2,\ldots,a_s$ 个,$a_1+a_2+\ldots+a_s=m$,未被更新的类为 $C_1,C_2,\ldots,C_r$,$r+s=k$,k 为类的总数量,则信息熵的变化如下:

$$Ent(S_l)_{new}=-\left(\sum_{i=1}^{r}\frac{|C_i|}{|S_l|+m}\log\left(\frac{|C_i|}{|S_l|+m}\right)+\sum_{j=1}^{s}\frac{|\bar{C}_j|+a_j}{|S_l|+m}\log\left(\frac{|\bar{C}_j|+a_j}{|S_l|+m}\right)\right).$$

类似信息熵增量公式的推导过程,$Ent(S_l)_{new}$–$Ent(S_l)$后可得到分母与分子分别为

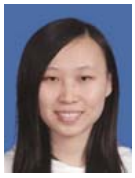
$$\left.\begin{aligned}
&den\left(Ent\left(S_l\right)_{new}-Ent\left(S_l\right)\right)=-\left(\left|S_l\right|+m\right)\left|S_l\right|num\left(Ent\left(S_l\right)_{new}-Ent\left(S_l\right)\right)\\
&=\left[\sum_{i=1}^{r}\left|S_l\right|\left|C_i\right|\log\left(\left|C_i\right|\right)+\sum_{j=1}^{s}\left|S_l\right|\left|\bar{C}_j\right|\log\log\left(\left\|\bar{C}_j\right\|+a_j\right)\right.\\
&\underline{\left.-m\sum_{i=1}^{k}\left|S_l\right|\left|C_i\right|\log\left(\left|S_l\right|+m\right)}+\left|S_l\right|\log\left(\left|C_j\right|+a_j\right)-\left|S_l\right|\log\left(\left|S_l\right|+m\right)\right]\\
&-\left[\sum_{i=1}^{r}\left|S_l\right|\left|C_i\right|\log\left(\left|C_i\right|\right)+\sum_{j=1}^{s}\left|S_l\right|\left\|\bar{C}_j\right\|\log\left(\left\|\bar{C}_j\right\|\right)-\underline{m\sum_{i=1}^{k}\left|S_l\right|\left|C_i\right|\log\left(\left|S_l\right|\right)}\right.\\
&\left.-m\sum_{i=1}^{k}\left(\left|C_i\right|\log\left(\left|C_i\right|\right)-\left|C_i\right|\log\left(\left|S_l\right|\right)\right)\right]
\end{aligned}\right\}\tag{10}$$

同理,式(10)中灰色部分可抵消,将下划线部分合并,整理余项并将分母分子合并后就可得到:

$$Ent\left(S_l\right)_{new}-Ent\left(S_l\right)=\frac{M}{\left|S_l\right|+m},$$

其中,

$$M=\left|S_l\right|\cdot\log\left(\frac{\left|S_l\right|+m}{\left|S_l\right|}\right)+mEnt\left(S_l\right)-\sum_{j=1}^{s}\left(\left|\bar{C}_j\right|\cdot\log\left(\frac{\left|\bar{C}_j\right|+a_j}{\left|\bar{C}_j\right|}\right)+a_j\log\left(\frac{\left|\bar{C}_j\right|+a_j}{\left|S_l\right|+m}\right)\right).$$



吴珉(1986－),女,云南昆明人,硕士,主要研究领域为轨迹数据管理.

唐祖锴(1977－),男,博士,副教授,CCF 专业会员,主要研究领域为软件工程,物联网工程,数据库技术.

祝园园(1984－),女,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘.

彭煜玮(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为时空数据管理,数据库管理系统,高端制造业大数据管理.

彭智勇(1963－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为复杂数据管理,可信数据管理,Web 数据管理.