Instituto Federal do Rio Grande do Norte
Campus Natal - Central
Diretoria de Gestão e Tecnologia da Informação
Tecnologia em Análise e Desenvolvimento de Sistemas

An Spatial-Temporal Model to Explore Interesting Dense Regions over Time

Felipe Mateus Freire Pontes

Natal-RN
Mês (por extenso) e ano

Felipe Mateus Freire Pontes

An Spatial-Temporal Model to Explore Interesting Dense Regions over Time

Trabalho de conclusão de curso de graduação do curso de Tecnologia e Análise em Desenvolvimento de Sistemas da Diretoria de Gestão e Tecnologia de Informação do Instituto Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Tecnologo em Análise e Desenvolvimento de Sistemas.

Linha de pesquisa:
Nome da linha de pesquisa

Orientador

Dr. Plácido Antônio de Souza Neto

TADS – Curso de Tecnologia em Análise e Desenvolvimento de Sistemas
DIATINF – Diretoria Acadêmica de gestão e Tecnologia da Informação
CNAT – Campus Natal - Central
IFRN – Instituto Federal do Rio Grande do Norte

Natal-RN

Mês e ano

Trabalho de Conclusão de Curso de Graduação sob o título *An Spatial-Temporal Model to Explore Interesting Dense Regions over Time* apresentada por Felipe Mateus Freire Pontes e aceita pelo Diretoria de Gestão e Tecnologia da Informação do Instituto Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Dr. Plácido Antônio de Souza Neto
Presidente
DIATINF – Diretoria Acadêmica de Gestão e Tecnologia da Informação
IFRN – Instituto Federal do Rio Grande do Norte

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Instituto

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Universidade

Natal-RN, data da defesa (dia, mês e ano).

Homenagem que o autor presta a uma ou mais pessoas.

Agradecimentos

Agradecimentos dirigidos àqueles que contribuíram de maneira relevante à elaboração do trabalho, sejam eles pessoas ou mesmo organizações.

Citação

Autor

An Spatial-Temporal Model to Explore Interesting Dense Regions over Time

Author: Felipe Mateus Freire Pontes
Supervisor: Dr. Plácido Antônio de Souza Neto

Abstract

O resumo em língua estrangeira (em inglês *Abstract*, em espanhol *Resumen*, em francês *Résumé*) é uma versão do resumo escrito na língua vernícula para idioma de divulgação internacional. Ele deve apresentar as mesmas características do anterior (incluindo as mesmas palavras, isto é, seu conteúdo não deve diferir do resumo anterior), bem como ser seguido das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores, na língua estrangeira. Embora a especificação abaixo considere o inglês como língua estrangeira (o mais comum), não fica impedido a adoção de outras linguas (a exemplo de espanhol ou francês) para redação do resumo em língua estrangeira.

Keywords: Keyword 1, Keyword 2, Keyword 3.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

Sumário

1 Introdução

Mais do que nunca estamos sobrecarregados com a quantidade de dados que criamos a cada dia. Quando comparamos quanto de informação vem sendo gerada nos ultimos anos, percebemos que está aumentando significamente. Além dessa evolução quantitativa, hoje temos os mais diversos tipos de informação, por exemplo: documentos, tuítes, fotos, vídeos, *GIFs*, *check-ins* entre vários outros.

Esse fenômeno vem sido chamado de *Big Data* e representa uma crescente área de estudo atualmente. Como consequência, pesquisadores estão analisando e aprendendo com essas informações geradas, entretanto o crescimento contínuo da quantidade de dados dificulta as análises. Portanto pessoas estão investindo em novas técnicas e ferramentas para romper desafios como mineração de dados, *data cleaning*, visualização de dados, classificação de dados, exploração de dados e muito mais.

Um tipo comum de dados é o que chamamos de dado espacial, o qual a informação possui atributos geográficos como latitude e longitudade (por exemplo: tuítes, avaliação de restaurantes, *check-ins* em estabelecimentos). Dados espaciais podem ser muito significativo, por exemplo, um *check-in* no aeroporto por sua irmã na manhã do seu aniversário, provavelmente significa que você terá uma surpresa.

Cada registro em dados espaciais representa uma atividade numa precisa localização geográfica, em outras palavras, a análise desse tipo de dado permite realizar descobertas baseadas em fatos. Analistas estão frequetemente interessados em observar padrões espaciais e tendências para melhorar seus processos de tomada de decisão. Análise de dados espaciais tem várias aplicações como gerenciamento de cidade inteligentes, gerenciamento de disastres e transporte autônomo (RODDICK et al., 2004; TELANG; PADMANABHAN; DESHPANDE, 2012).

1.1 Problema

A análise de dados espaciais geralmente é realizada num contexto exploratório: o analista não tem uma consulta precisa em mente e ele explora os dados em passos iterativos a fim de encontrar resultados potencialmente interessantes. Tradicionalmente, um cenário de análise exploratória é descrito na seguinte maneira: o analista visualiza um subconjunto de dados usando uma consulta em ambiente de visualização (por exemplo: Tableau[1], Exhibit[2], Spotfire[3]); o resultado será ilustrado em um mapa geográfico; então o analista investiga diferentes partes do conjuto de dados movendo ou focando o mapa afim de encontrar padrões ou tendências de interesse. O analista pode iterar por esse processo várias vezes realizando consultas diferentes e focando em diferentes aspectos.

Contudo, a vasto tamanho do conjunto de dados espacias faz com que o analista se sinta perdido durante a exploração. É possível ter milhares de pontos geográficos em cada bairro de uma cidade, por exemplo. Analistas precisam ter acesso apenas a algumas opções (chamadas de "highlights") que ajam como uma direção e assim permitir que ele foque no que lhe interessa na análise. No cenário perfeito, essas opções não são aleatoriamente escolhidas e representam o que o analista se mostrou interessado em iterações passadas.

Neste trabalho, formulamos uma solução para "realçamento de dados usando feedback coletado ao longo do tempo". Em outras palavras, buscamos realçar alguns pontos geográficos baseado nos interesses do analista afim de guiá-lo na direção ao que ele deve se concentrar nas iterações seguintes do processo de análise.

1.1.1 Case Study

Now, we will present a case study in order to show the functionality of our approach in practice.

Example. *Lucas is planning to spend few days in Paris, France. His appreciation of French culture makes him interested in new experiences in the city. He decides to rent a home-stay from Airbnb website*[4]*. He likes to discover the city, hence he is open to any type of lodging in any region with an interest to stay in the city center. The website returns* 4000 *different locations. As he has no other preferences, an exhaustive investigation needs scanning each location independently which is nearly infeasible. While he is scanning few*

[1] *http://www.tableau.com*
[2] *http://www.simile-widgets.org/exhibit/*
[3] *http://spotfire.tibco.com*
[4] *http://www.airbnb.com*

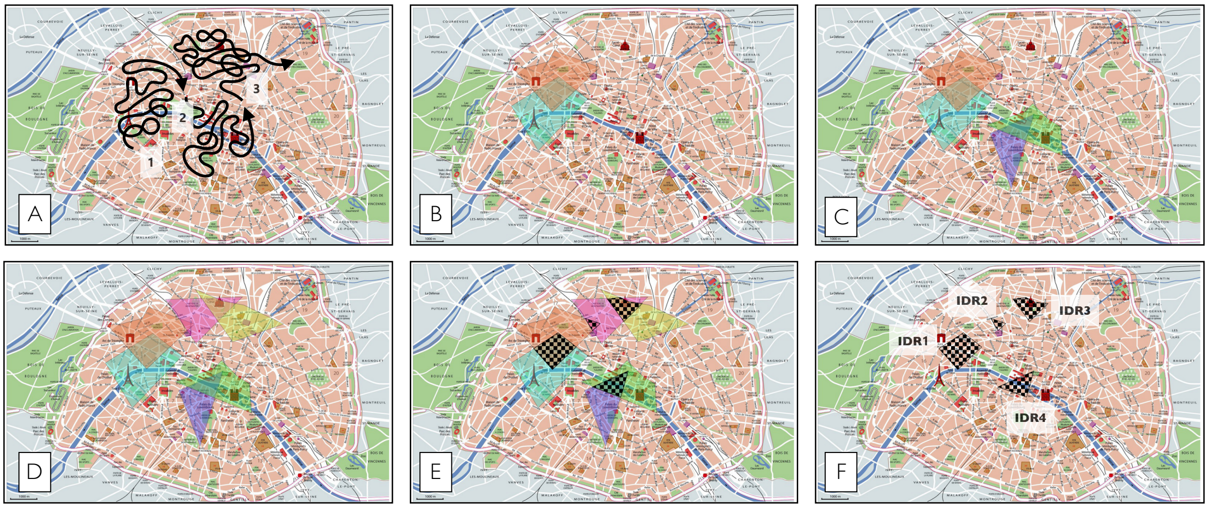

Figura 1: The process of exploring Paris home-stays.

first options, he shows interest in the region of "Champ de Mars" (near Eiffel Tower), but he forgets or doesn't feel necessary to click a point there. By collecting feedback on his mouse moves over the home-stays in Paris, our system can quickly detect his interest in the region and short-list a small subset of locations (i.e., highlights) accordingly to be recommended to Lucas.

We follow the above example to describe how implicit feedback is collected in action. Figure 1 shows Lucas' steps to explore home-stays in Paris. Figure 1.A shows his mouse movements in different time stages. In this example, we consider $g = 3$ and capture Lucas' feedback in three different time segments (progressing from Figures 1.B to 1.D). It shows that Lucas started his search around Eiffel Tower and Arc de Triomphe (Figure 1.B) and gradually showed interest in south (Figure 1.C) and north (Figure 1.D) as well. All intersections between those regions are discovered (hatching regions in Figure 1.E) which will constitute the set of *Interesting Dense Regions* (Figure 1.F), i.e., IDR1 to IDR4.

What if Lucas wanted to come back to Paris, France next year? He will have to repeat the same exploratory analyse, unless he remember the exact location of home-stays he showed interest last year. Using our system, he won't need to remember, because his preferences were collected and can be used to highlight a subset of similar home-stays.

In the context of exploratory analysis, the analyst may change his preferences between session (e.g., in the winter, Lucas may want to be close to the Eiffel Tower, but in the summer, he may not). In order to tackle this challenge we also apply a temporal analyse to identify patterns in how the analyst preferences change between sessions which allow our highlighting method to be more precise and consistent to the analyst interest.

1.2 Objectives

In this section, we define the general and specific objectives of our work.

1.2.1 General Objectives

- Introduce a time-aware guidance approach for spatial data exploration;
- Elaborate how temporal analyses can be effectively applied in data exploration;

1.2.2 Specific Objectives

- Describe our data model used for temporal analyses;
- Describe our concept of *Interesting Dense Regions* used for collecting feedback;
- Present the results of our guidance approach.

1.3 Organization

The next chapters is as follow: in the Chapter 2 we discuss the background of this work. Chapter 3 defines the data model. Chapter 4 presents how the feedback is collected during exploration. Chapter 5 presents how temporal analysis is applied. Chapter 6 presents how highlight interesting points in order to guide the user using collected feedback and results from temporal analysis. Chapter 7 shows experiments and its results. Chapter 8 presents some conclusions and future directions.

2 Background

This chapter gives an overview of related work in literature about feedback exploitation, information-highlighting methods and temporal analysis applications. We also present the system we are extending.

2.1 Related Work

The literature in spatial data analysis has a focus on *efficiency* of exploratory interactions. The common approach is to design pre-computed index which enable efficient retrieval of spatial data (LINS; KLOSOWSKI; SCHEIDEGGER, 2013). However, we should also put attention in the *value* of spatial data, because is very common to see an analyst getting lost in the huge amount of geographical points. In order to overcome this challenge, visualization environments (e.g., Tableau[1], Exhibit[2], Spotfire[3]) offer features to manipulate data (e.g., filters, aggregate queries, etc).

2.1.1 Feedback exploitation

Our proposed spatial-temporal model leverage the spatial data analysis by exploiting collected feedback during the analyst exploration to highlight subsets of geographical points. In the literature, are several instances of feedback exploitation to guide the analysts in further analysis steps (e.g., Boley et al. (2013)). The common approach is a top-k processing methodology in order to prune the search space based on the explicit feedback and recommend a small subset of interesting results of size k. A clear distinction of our work is that it doesn't aim for pruning, but leveraging the actual data with potential interesting results that the analyst may miss due to the huge volume of spatial data. While in top-k processing algorithms, analyst choices are limited to k, we offer the freedom of

[1] *http://www.tableau.com*
[2] *http://www.simile-widgets.org/exhibit/*
[3] *http://spotfire.tibco.com*

choice where highlights get seamlessly updated with new analyst choices.

2.1.2 Information-highlighting methods

There exist few instances of information-highlighting methods in the literature: Liang e Huang (2010), Robinson (2011), Wongsuphasawat et al. (2016), Willett, Heer e Agrawala (2007). All these methods are *objective* and do not apply to the context of spatial guidance where user feedback is involved. In terms of recommendation, few approaches focus on spatial dimension (BAO et al., 2015; LEVANDOSKI et al., 2012) while the context and result diversification are missing.

2.1.3 Temporal analysis applications

There are currently several instances which combine temporal analysis with spatial data in the literature (e.g., Baculo et al. (2017), Balahadia e Trillanes (2017), Chidean et al. (2018), Ghahramani, Zhou e Hon (2018), Kamath e Caverlee (2013), Lopes-Teixeira, Batista e Ribeiro (2018), Ma et al. (2017), Mijović et al. (2016), Tomoki e Keiji (2010), Nara e Torrens (2007), Zhan et al. (2017), Zheng et al. (2018)). Those are applications of temporal analysis in specific context, which does not involve user feedback, but represent how temporal analysis could be insightful.

Baculo et al. (2017) and Balahadia e Trillanes (2017) make use of public data of Manila, the most densely populated city in the Philippines, to combine spatial data, temporal analysis and prediction model to allow decision makers to prepare an effective public management plan. Ma et al. (2017) and Zheng et al. (2018) also perform real-world analyses of how events (e.g., protests) impact the taxi trajectories which results could provide helpful insights for traffic control and transit service plans for city administrators. Both perform insightful analyses which we will use as inspiration.

Chidean et al. (2018) present how to detect spatial-temporal patterns in the context of wind power resource in the Iberian Peninsula using Second-Order Data-Coupled Clustering algorithm. Despite the detailed study, it does not work in a exploratory context.

Ghahramani, Zhou e Hon (2018), Lopes-Teixeira, Batista e Ribeiro (2018) and Zhan et al. (2017) demostrate how temporal analyses can be applied in the geographical context. Zhan et al. (2017) goes deeper generating a hierarchical cluster tree. Regardless of insights and methods, it does not contribute to the subject in question.

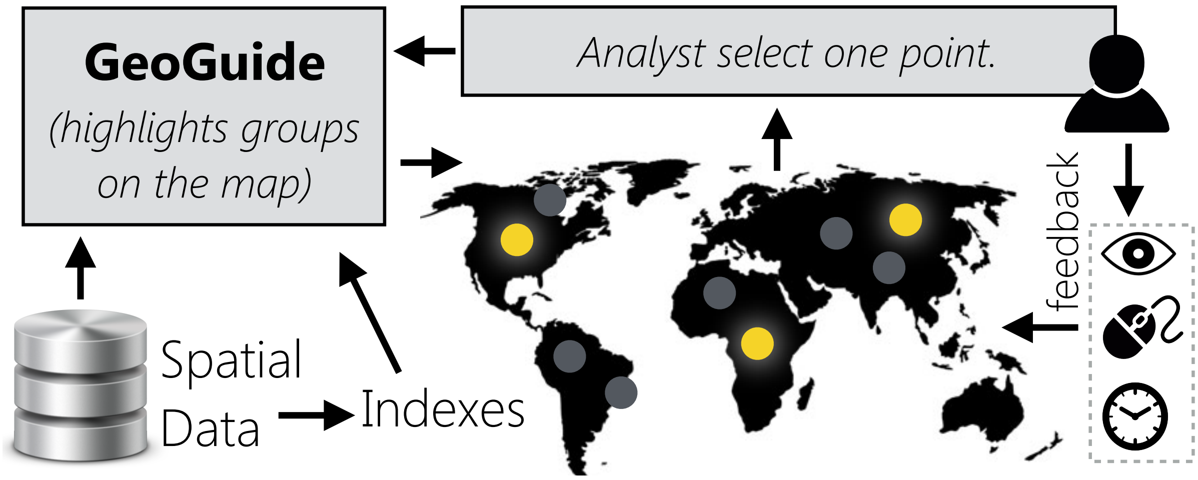

Figura 2: GeoGuide Framework

Kamath e Caverlee (2013) propose a novel reinforcement learning approach to predict events (i.e., online meme) in the spatial-temporal context.

Nara e Torrens (2007) introduce a 3D visualization of space-time which helps to qualitatively and quantitatively analyze the spatiotemporal patterns and tendencies. We will make use of this visualization approach to display our collected data in chapter 4.

2.2 GeoGuide

GeoGuide (OMIDVAR-TEHRANI et al., 2017) is a spatial data visualization environment which keep track of user preferences during exploration in order to use collected feedback to highlight subsets of geographical points that may be interesting to the analyst. Figure 2 illustrates the main components of GeoGuide architecture which we will present in the next subsections.

2.2.1 Preprocessing

GeoGuide requires a preprocessing step in order to create a index which will be used during highlighting. The index is a comparative table between every points with two quality metrics, i.e., relevance and diversity.

2.2.1.1 Relevance

Relevance represent how a point a is similar to a point b in the current dataset. GeoGuide use the relevance to highlight points in the same line with the analyst feedback.

2.2.1.2 Diversity

Diversity represent how distant is the region where a point a is to the region where point b is located. It allows the analyst to explore different regions, but still work with relevant points to his interest.

2.2.2 Tracking User Preferences

In order to keep track of user preferences, GeoGuide use both explicit and implicit feedback. Explicit feedback is when the user is analyzing the attributes of a point (e.g., the house description in a Airbnb context) and explicitly ask to explore similar points to the current selected one. Implicit feedback is tracked using the mouse movements, gaze tracking and metrics like "how long the user was analyzing the profile of a point".

2.2.3 Highlighting Spatial Data

GeoGuide combine both preprocessed index and collected feedback to highlight a subset of spatial data according to the analyst preferences. GeoGuide highlighting feature prove to be efficient in terms of "how many steps the analyst takes until complete a task of finding a point in a request location". Using GeoGuide the analysts were able to complete the task using in average 10.7 steps, while using Tableau, they took about 43 steps.

In this work, we will leverage GeoGuide into two new concepts: i. interesting dense regions and ii. understanding how the user preferences change over time.

3 Data Model Definition

3.1 Spatial layer

Each point in a dataset ($p \in \mathcal{P}$) is described using its coordinates (latitude and longitude) and also associated with a set of attributes ($dom(p)$). For instance, TODO

3.2 Interesting Dense Regions

TODO

We have IDRs per iteration/session where implicit feedback is captured such mouse moves (or eye gaze). In the beginning, each IDRs is a group of raw points described using its coordinates (latitude and longitude) and a timestamp (the unix timestamp it was captured). These raw points once captured will enter the clustering (for now, ST-DBSCAN) phase to generate the IDR itself with a profile. The profile is built based on the spatial layer and it should represent a summary of its contained points from the spatial layer.

- A profile has summary of its spatial points number attributes. For each number attribute in $dom(p)$, we calculate the average, median and standard deviation based on the points contained in the IDR.
- A profile has a word rank R of the terms in the text attributes of its spatial points. For each text attribute in $dom(p)$, we evaluate the most used terms in order to create a word rank (KUMAR; KAUR, 2017).
- A profile has a map M between the $< name, value >$ of categoricals attributes and its relevance in $dom(p)$.
- TODO: datetime attributes
- A profile has a meta property with values such the count of points in the IDR.

3.3 Highlighting

TODO

4 Collecting feedback

TODO

5 Applying temporal analysis

TODO

6 Guiding the user

TODO

7 Experiments

TODO

7.1 Results

TODO

8 Conclusion

To our knowledge...

8.1 Contributions

TODO

8.2 Restrictions

TODO

8.3 Future work

TODO

Referências

BACULO, M. J. C. et al. Geospatial-temporal analysis and classification of criminal data in manila. In: *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. [S.l.: s.n.], 2017. p. 6–11.

BALAHADIA, F. F.; TRILLANES, A. O. Improving fire services using spatio-temporal analysis: Fire incidents in manila. In: *2017 IEEE Region 10 Symposium (TENSYMP)*. [S.l.: s.n.], 2017. p. 1–5.

BAO, J. et al. Recommendations in location-based social networks: a survey. *GeoInformatica*, v. 19, n. 3, p. 525–565, 2015. Disponível em: <http://dx.doi.org/10.1007/s10707-014-0220-8>.

BOLEY, M. et al. One click mining: Interactive local pattern discovery through implicit preference and performance learning. In: ACM. *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*. [S.l.], 2013. p. 27–35.

CHIDEAN, M. I. et al. Spatio-temporal analysis of wind resource in the iberian peninsula with data-coupled clustering. *Renewable and Sustainable Energy Reviews*, v. 81, p. 2684 – 2694, 2018. ISSN 1364-0321. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1364032117310080>.

GHAHRAMANI, M.; ZHOU, M.; HON, C. T. Spatio-temporal analysis of mobile phone data for interaction recognition. In: *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. [S.l.: s.n.], 2018. p. 1–6.

KAMATH, K. Y.; CAVERLEE, J. Spatio-temporal meme prediction: Learning what hashtags will be popular where. In: *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*. New York, NY, USA: ACM, 2013. (CIKM '13), p. 1341–1350. ISBN 978-1-4503-2263-8. Disponível em: <http://doi.acm.org/10.1145/2505515.2505579>.

KUMAR, H.; KAUR, H. Clustering and ranking social media users based on temporal analysis. In: *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*. [S.l.: s.n.], 2017. p. 271–275.

LEVANDOSKI, J. J. et al. Lars: A location-aware recommender system. In: *ICDE*. [s.n.], 2012. p. 450–461. ISBN 978-0-7695-4747-3. Disponível em: <http://dx.doi.org/10.1109/ICDE.2012.54>.

LIANG, J.; HUANG, M. L. Highlighting in information visualization: A survey. In: *2010 14th International Conference Information Visualisation*. [S.l.: s.n.], 2010. ISSN 1550-6037.

LINS, L.; KLOSOWSKI, J. T.; SCHEIDEGGER, C. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 19, n. 12, p. 2456–2465, 2013.

LOPES-TEIXEIRA, D.; BATISTA, F.; RIBEIRO, R. Spatio-temporal analysis of brand interest using social networks. In: *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2018. p. 1–6.

MA, J. W. et al. Spatio-temporal factor analysis of characterizing mass protest events using taxi trajectory in seoul, korea. In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Analytics for Local Events and News*. New York, NY, USA: ACM, 2017. (LENS'17), p. 6:1–6:7. ISBN 978-1-4503-5500-1. Disponível em: <http://doi.acm.org/10.1145/3148044.3148050>.

MIJOVIć, V. et al. Exploratory spatio-temporal analysis of linked statistical data. *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 41, p. 1 – 8, 2016. ISSN 1570-8268. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826816300488>.

NARA, A.; TORRENS, P. M. Spatial and temporal analysis of pedestrian egress behavior and efficiency. In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 2007. (GIS '07), p. 59:1–59:4. ISBN 978-1-59593-914-2. Disponível em: <http://doi.acm.org/10.1145/1341012.1341083>.

OMIDVAR-TEHRANI, B. et al. Geoguide: An interactive guidance approach for spatial data. In: *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, United Kingdom, June 21-23, 2017*. [s.n.], 2017. p. 1112–1117. Disponível em: <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.170>.

ROBINSON, A. C. Highlighting in geovisualization. *Cartography and Geographic Information Science*, v. 38, n. 4, p. 373–383, 2011. Disponível em: <http://dx.doi.org/10.1559/15230406384373>.

RODDICK, J. F. et al. Spatial, temporal and spatio-temporal databases - hot issues and directions for phd research. *SIGMOD Record*, v. 33, n. 2, p. 126–131, 2004. Disponível em: <http://doi.acm.org/10.1145/1024694.1024724>.

TELANG, A.; PADMANABHAN, D.; DESHPANDE, P. Spatio-temporal indexing: Current scenario, challenges and approaches. In: *Proceedings of the 18th International Conference on Management of Data*. Mumbai, India, India: Computer Society of India, 2012. (COMAD '12), p. 9–11. Disponível em: <http://dl.acm.org/citation.cfm?id=2694443.2694449>.

TOMOKI, N.; KEIJI, Y. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, v. 14, n. 3, p. 223–239, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2010.01194.x>.

WILLETT, W.; HEER, J.; AGRAWALA, M. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 13, n. 6, p. 1129–1136, 2007.

WONGSUPHASAWAT, K. et al. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *TVCG*, IEEE, v. 22, n. 1, 2016.

ZHAN, X. et al. Spatial-temporal analysis on bird habitat discovery in china. In: *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*. [S.l.: s.n.], 2017. p. 573–578.

ZHENG, L. et al. Spatial–temporal travel pattern mining using massive taxi trajectory data. *Physica A: Statistical Mechanics and its Applications*, v. 501, p. 24 – 41, 2018. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437118301419>.