

INSTITUTO FEDERAL DO RIO GRANDE DO NORTE
CAMPUS NATAL - CENTRAL
DIRETORIA DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Um Modelo Espaço-Temporal para Explorar Regiões Densas Interessantes

Felipe Mateus Freire Pontes

Natal-RN
Dezembro, 2018

Felipe Mateus Freire Pontes

Um Modelo Espaço-Temporal para Explorar Regiões Densas Interessantes

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 Tecnólogo em Análise e Desenvolvimento de Sistemas.

Linha de pesquisa:
Banco de Dados

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

Dezembro, 2018

Trabalho de Conclusão de Curso de Graduação sob o título *Um Modelo Espaço-Temporal para Explorar Regiões Densas Interessantes* 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, 06 de Dezembro de 2018.

Aos meus pais que nunca duvidaram de mim.

Agradecimentos

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

A coisa mais autêntica sobre nós é nossa capacidade de criar, de superar, de suportar, de transformar, de amar e de sermos maiores que nosso sofrimento.

Ben Okri

Um Modelo Espaço-Temporal para Explorar Regiões Densas Interessantes

Autor: Felipe Mateus Freire Pontes

Orientador(a): Dr. Plácido Antônio de Souza Neto

RESUMO

O resumo deve apresentar de forma concisa os pontos relevantes de um texto, fornecendo uma visão rápida e clara do conteúdo e das conclusões do trabalho. O texto, redigido na forma impersonal do verbo, é constituído de uma sequência de frases concisas e objetivas e não de uma simples enumeração de tópicos, não ultrapassando 500 palavras, seguido, logo abaixo, das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores. Por fim, deve-se evitar, na redação do resumo, o uso de parágrafos (em geral resumos são escritos em parágrafo único), bem como de fórmulas, diagramas e símbolos, optando-se, quando necessário, pela transcrição na forma extensa, além de não incluir citações bibliográficas.

Palavras-chave: Palavra-chave 1, Palavra-chave 2, Palavra-chave 3.

An Spatial-Temporal Model for Exploring Interesting Dense Regions

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 vernácula 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 línguas (a exemplo de espanhol ou francês) para redação do resumo em língua estrangeira.

Keywords: Keyword 1, Keyword 2, Keyword 3.

Lista de figuras

1	O processo de explorar estadias em Paris.	p. 16
2	Componentes do GeoGuide	p. 21
3	Exemplo de pontos geográficos com seus atributos	p. 22
4	Processo de Descoberta de IDRs.	p. 26
5	Evolução no contexto espacial.	p. 30

Lista de tabelas

1	Exemplo de Índice de Relevância e Diversidade para pontos na Figura 3	p. 23
2	Exemplo de atributos de um ponto que representa uma estadia	p. 27
3	Exemplo de perfil dos atributos numéricos de uma IDR	p. 28
4	Exemplo parcial de perfil do atributo textual de uma IDR	p. 28
5	Exemplo de perfil dos atributos categóricos de uma IDR	p. 28

Lista de abreviaturas e siglas

IDR Interesting Dense Region

Lista de Algoritmos

1 Descobrimento de IDRs p. 25

Sumário

1	Introdução	p. 14
1.1	Problema	p. 15
1.1.1	Caso de Estudo	p. 15
1.2	Objetivos	p. 17
1.2.1	Objetivo Geral	p. 17
1.2.2	Objetivos Específicos	p. 17
1.3	Organização	p. 17
2	Contextualização	p. 19
2.1	Trabalhos Relacionados	p. 19
2.1.1	Exploração de Feedback	p. 19
2.1.2	Métodos de Destacamento de Informação	p. 20
2.1.3	Aplicações de Análise Temporal	p. 20
2.2	GeoGuide	p. 21
2.2.1	Preprocessamento	p. 22
2.2.1.1	Relevância e Diversidade	p. 22
2.2.2	Preferências do Usuário	p. 23
2.2.3	Destacamento de Dados Espaciais	p. 23
3	Modelo Espaço-Temporal	p. 24
3.1	Regiões Densas Interessantes	p. 24
3.1.1	Perfil	p. 26

3.2	Modelo Temporal	p. 29
3.3	Contexto Espacial	p. 30
3.4	Contexto de Domínio	p. 30
4	Conclusão	p. 31
4.1	Contribuições	p. 31
4.2	Trabalhos futuros	p. 31
	Referências	p. 32

1 Introdução

Mais do que nunca estamos sobrecarregados com a quantidade de dados que criamos a cada dia (PRADEEP; KALLIMANI, 2017). Quando comparamos quanto de informação vem sendo gerada nos últimos 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 (ZHANG et al., 2015).

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

Cada registro de 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 frequentemente 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 desastres 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: Tableau¹, Exhibit², Spotfire³); o resultado será ilustrado em um mapa geográfico; então o analista investiga diferentes partes do conjunto de dados movendo ou focando uma região do 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 espaciais 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 um modelo para permitir o “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 Caso de Estudo

Nessa seção, vamos apresentar um caso de estudo afim de demostrar a funcionalidade da nossa abordagem na prática.

Exemplo. *Lucas está planejando passar alguns dias em Paris, França. Sua apreciação pela cultura francesa faz como que ele tenha interesse em novas experiências na cidade. Ele decidiu por alugar uma estadia pelo Airbnb⁴. Ele gosta de descobrir a cidade, portanto ele é aberto a qualquer tipo de estadia em qualquer região com um leve interesse em ficar perto do centro da cidade. O sistema retorna 4000 opções diferentes. Como ele não tem outras preferências, uma investigação exaustiva para avaliar cada região da cidade*

¹<http://www.tableau.com>

²<http://www.simile-widgets.org/exhibit/>

³<http://spotfire.tibco.com>

⁴<http://www.airbnb.com>

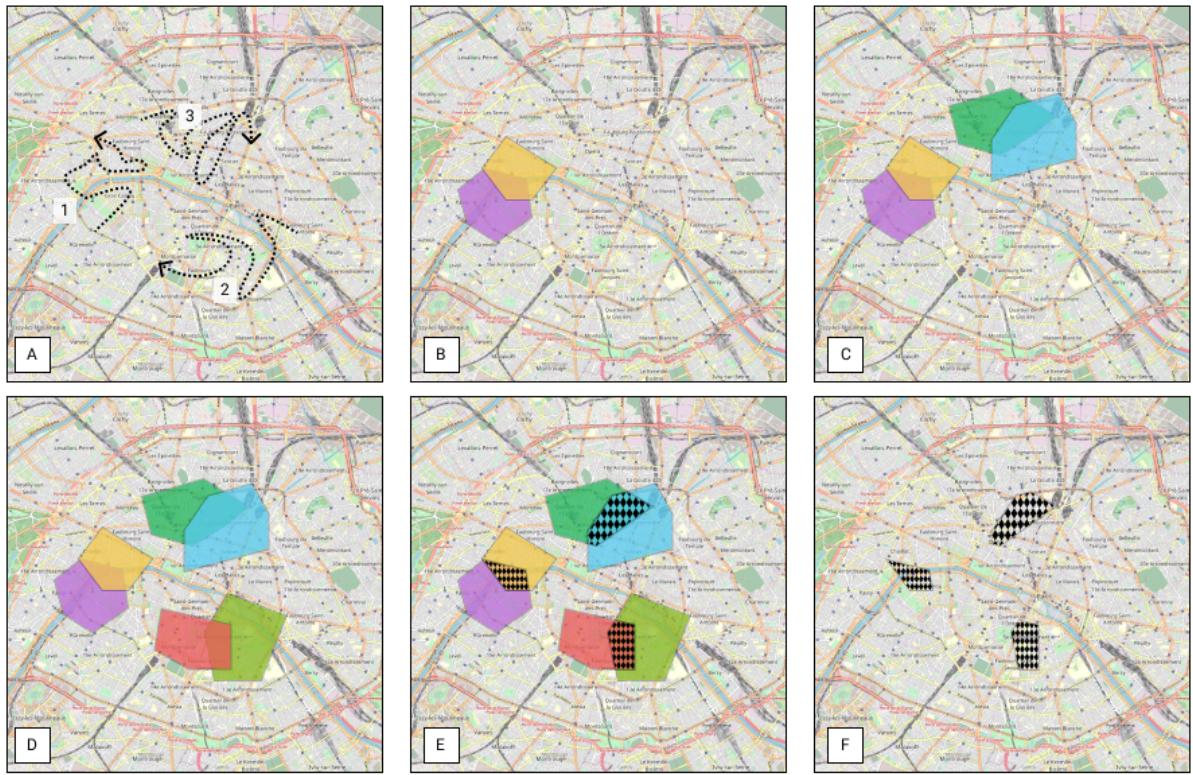


Figura 1: O processo de explorar estadias em Paris.

independentemente é necessário, o que é quase impossível. Enquanto estava avaliando algumas opções, ele demonstrou interesse na região de “Champ de Mars” (próximo à Torre Eiffel), mas ele esqueceu ou não achou necessário clicar num ponto nessa região. Coletando o feedback do seus movimentos com o mouse no mapa de estadias em Paris, nosso sistema consegue de maneira transparente detectar o interesse dele na região supracitada e apresentar uma quantidade pequena de opções recomendadas para Lucas.

Seguimos o exemplo acima para descrever como feedback implícito é coletado na prática. Imagem 1 mostra os passos de Lucas para explorar estadias em Paris. Imagem 1.A mostra os movimentos do mouse dele em diferentes intervalos de tempo. Nesse exemplo, coletamos o feedback de Lucas em 3 diferentes intervalos de tempo (evoluindo das Imagens 1.B até 1.D). Isso mostra que Lucas começou sua busca perto da Torre Eiffel e *Arc de Triomphe* (Imagen 1.B) e gradualmente mostrou também interesse no sul (Imagen 1.C) e norte (Imagen 1.D). Todas as interseções entre essas regiões são descobertas (regiões tachadas na Imagem 1.E), o que representa um conjunto de regiões onde o interesse de Lucas está direcionado e onde, provavelmente, ele vai decidir ficar durante sua visita à Paris.

E se Lucas quiser voltar para Paris próximo ano? Ele teria que repetir a mesma análise

exploratória, a não ser que ele lembre a localização exata das estadias interessantes à ele no ano passado. Usando nosso sistema, ele não precisaria lembrar, porque suas preferências foram coletadas e poderiam ser usadas para realçar um subconjunto similar ao do ano anterior.

No contexto da análise exploratória, o analista talvez mude suas preferências entre as sessões (por exemplo, no inverno, Lucas talvez queira ficar próximo ao Torre Eiffel, mas no verão, ele talvez queira ficar mais distante dos pontos turísticos). Afim de atacar esse desafio, nosso modelo permite uma análise temporal para identificar padrões em como as preferências dos analistas mudam entre as sessões o que permite nosso método de realçamento ser mais preciso e consistente com o interesse do analista mesmo em momentos diferentes do ano.

1.2 Objetivos

Nessa seção, definimos os objetivos gerais e específicos do nosso trabalho.

1.2.1 Objetivo Geral

- Propor um modelo espaço-temporal para orientação na exploração de dados espaciais.

1.2.2 Objetivos Específicos

- Descrever nosso modelo de dados usado para análise temporal;
- Descrever nosso conceito de Região Densas Interessantes usado para captura de feedback;
- Elaborar como análise temporal pode ser efetivamente aplicada na exploração de dados;
- Descrever como nosso modelo pode ser estendido à diversos contextos.

1.3 Organização

Os próximos capítulos estão organizados na seguinte maneira: no Capítulo 2 discutimos o estado da arte por trás desse trabalho; Capítulo 3 define o modelo de dados,

apresenta como é feito a coleta de feedback durante a análise exploratória, demonstra como a análise temporal pode ser aplicada. Capítulo 4 conclui e propõe futuros trabalhos.

2 Contextualização

Este capítulo dá uma visão geral sobre os trabalhos relacionados sobre exploração de feedback, métodos de destaqueamento de informações e aplicação de análises temporais. Também apresentamos o sistema que estamos estendendo neste trabalho.

2.1 Trabalhos Relacionados

A literatura na análise de dados espaciais possui um foco no eficiência das iterações exploratórias. A abordagem comum é projetar índices preprocessados, o quais permitem a consulta eficiente de dados espaciais (LINS; KLOSOWSKI; SCHEIDEGGER, 2013). No entanto, também devemos direcionar nossa atenção para o *valor* dos dados espaciais, porque é muito comum encontrar um analista se perdendo numa enorme quantidade de pontos geográficos. Para solucionar esse problema, ambientes de visualização, como, por exemplo, Tableau¹, Exhibit², Spotfire³, oferecem funcionalidades para manipular os dados como filtros, consultas agregadores entre outras. Entretanto essas funcionalidades não se mostram eficazes visto que nessas ferramentas o pesquisador precisa saber exatamente o que procura. Buscando otimizar essa análise exploratória combinamos a exploração de feedback, métodos de destaqueamento de informações e análise temporal neste trabalho.

2.1.1 Exploração de Feedback

Nosso modelo espaço-temporal aprimora o processo de análise de dados espaciais destacando subconjuntos de pontos geográficos com base no feedback coletado durante a exploração do analista. Na literatura, há várias trabalhos sobre exploração de feedback para orientar o analista nas futuras iterações da análise (por exemplo, Boley et al. (2013)). A abordagem comum é uma metodologia *top-k* para reduzir o escopo da consulta baseado

¹<http://www.tableau.com>

²<http://www.simile-widgets.org/exhibit/>

³<http://spotfire.tibco.com>

no feedback explícito e recomendar um pequeno subconjunto de resultados interessantes de tamanho k . Uma clara distinção do nosso trabalho é que não buscamos reduzir o escopo, mas alavancar o conjunto de dados com resultados potencialmente interessantes que o analista talvez não tenha notado devido ao enorme volume de dados espaciais. Enquanto as escolhas do analista são limitadas por k em algoritmos de *top-k* processamento, oferecemos a liberdade de escolha ao mesmo tempo que pontos geográficos vão sendo transparentemente destacados com base nas novas escolhas do analista.

2.1.2 Métodos de Destacamento de Informação

Há trabalhos na literatura sobre métodos de destacamento de informações, por exemplo: Liang e Huang (2010), Robinson (2011), Wongsuphasawat et al. (2016), Willett, Heer e Agrawala (2007). Entretanto todos esses métodos são *objetivos* e não são aplicáveis para o contexto de orientação espacial onde o feedback do usuário é envolvido. Em termos de recomendação, algumas abordagens focam na dimensão espacial (BAO et al., 2015; LEVANDOSKI et al., 2012) enquanto o contexto e a diversidade do resultado é deixado de lado.

2.1.3 Aplicações de Análise Temporal

Existem várias instâncias na literatura que combinam análise temporal com dados espaciais, como 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). Essas aplicações de análise temporal são em contextos específicos, os quais não envolve feedback do usuário, mas representam como análise temporal pode ser perspicaz e proveitosa.

Baculo et al. (2017) e Balahadia e Trillanes (2017) fazem uso de dados públicos de Manila, capital das Filipinas, combinando dados espaciais, análise temporal e modelos preditivos e mostrando resultados que podem ser utilizados para preparação de um plano de gestão pública eficaz. Ma et al. (2017) e Zheng et al. (2018) também fazem análises realistas de como eventos, como protestos, impactam na trajetórias de taxis, cujos resultados podem auxiliar no controle de tráfego urbano e nos planos de serviços de transporte da cidade. Ambos realizam ricas análises, as quais vamos usar como inspiração neste trabalho.

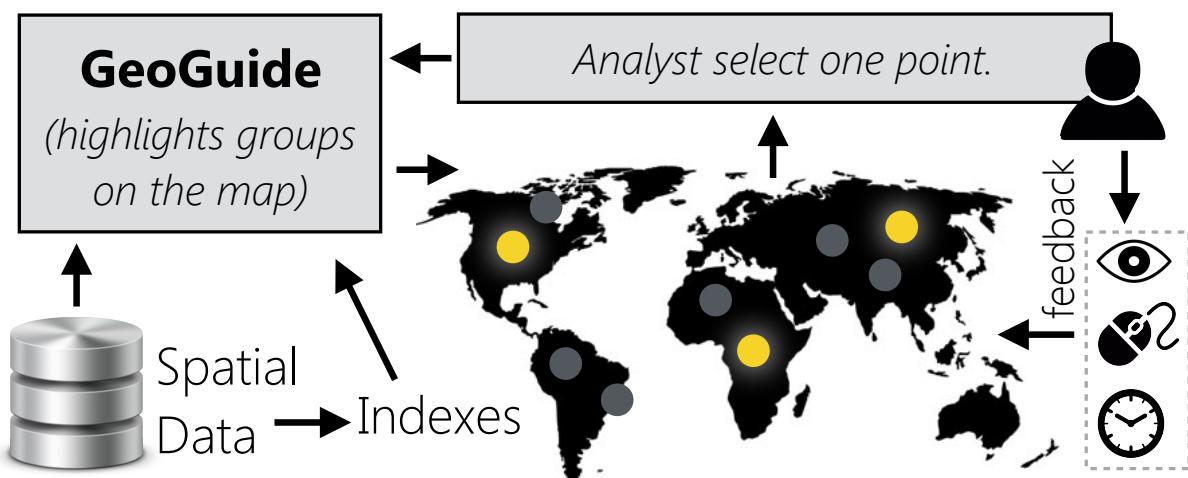


Figura 2: Componentes do GeoGuide

Chidean et al. (2018) apresenta como detectar padrões espaço-temporais no contexto do uso de energia eólica na Península Ibérica usando o algoritmo *Second-Order Data-Coupled Clustering*. Apesar do estudo detalhado, esse trabalho não contempla um contexto de análise exploratória.

Ghahramani, Zhou e Hon (2018), Lopes-Teixeira, Batista e Ribeiro (2018) e Zhan et al. (2017) demostram como análise temporal pode ser aplicado no contexto geográfico. Zhan et al. (2017) vai além gerando uma árvore de clusterização hierárquica. Apesar dos métodos e resultados serem bem detalhados nos trabalhos, essas contribuições não se aplicam para o assunto em questão.

Kamath e Caverlee (2013) propõe uma abordagem de *reinforcement learning* para prever eventos (adoção de *memes*) num contexto espaço-temporal. Nara e Torrens (2007) introduz um modelo de visualização 3D para dados espaço-temporais que ajuda a analisar qualitativa e quantitativamente os padrões e tendências espaço-temporais. Ambos os trabalhos representam como nosso modelo proposto pode ser combinado com diversas técnicas.

2.2 GeoGuide

GeoGuide (OMIDVAR-TEHRANI et al., 2017) é um ambiente de visualização de dados espaciais que coleta as preferências do usuário durante a exploração para destacar subconjuntos de pontos geográficos que podem ser interessantes ao analista. Figura 2 ilustra os principais componentes da arquitetura do GeoGuide que apresentaremos nas próximas subseções.

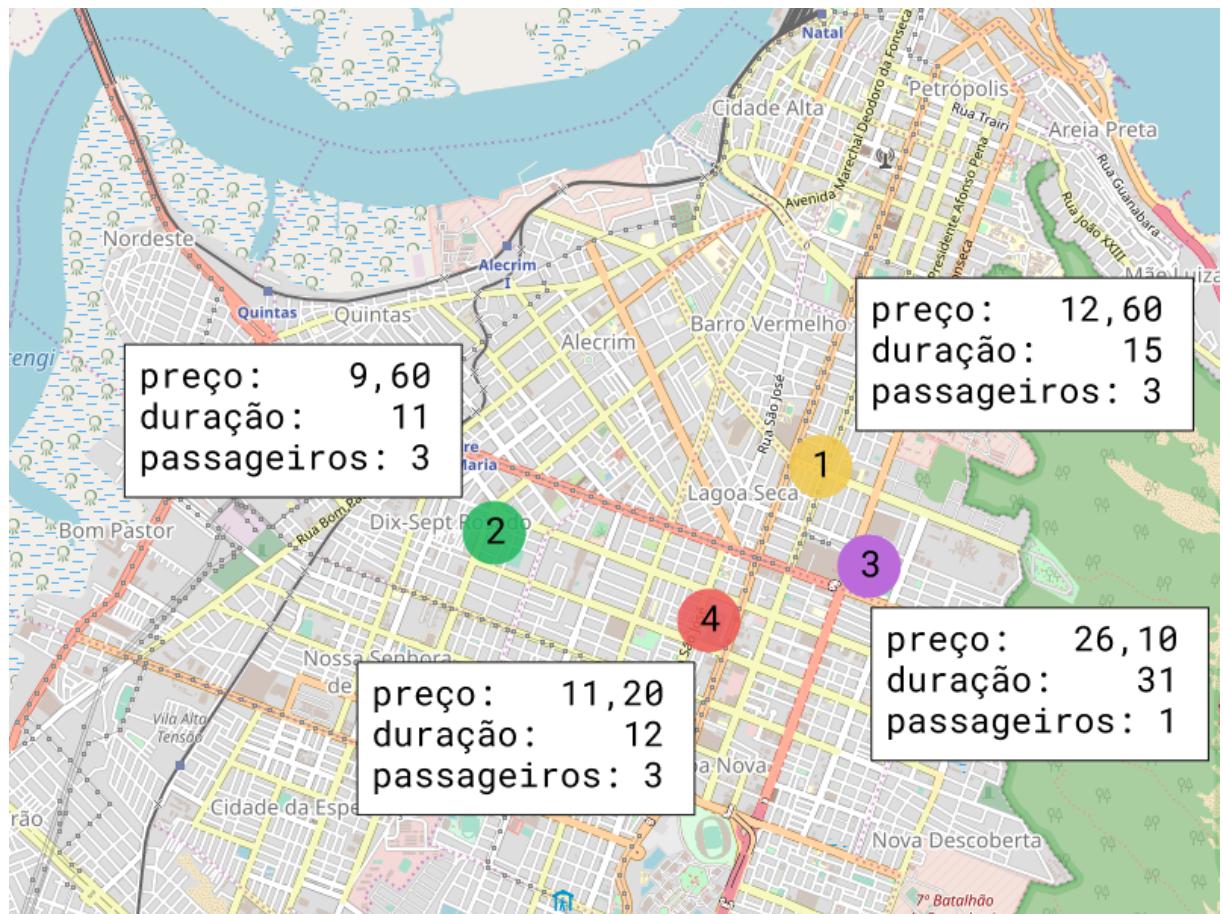


Figura 3: Exemplo de pontos geográficos com seus atributos

2.2.1 Preprocessamento

GeoGuide precisa de um passo de preprocessamento para criar os índices que serão usados durante a fase de destaque. O índice é uma tabela comparativa entre todos os pontos usando duas métricas de qualidade: relevância e diversidade. Os valores calculados são normalizados no intervalo de 0 à 1.

2.2.1.1 Relevância e Diversidade

Relevância representa o quanto similar é o ponto a com o ponto b num conjunto de dados. GeoGuide usa a relevância para destacar pontos similares ao feedback do analista. Diversidade representa quanto distante o ponto a está localizado do ponto b . GeoGuide usa a diversidade para permitir ao analista explorar diferentes regiões, mas ainda assim trabalhar com pontos relevantes ao seu interesse.

Na Figura 3, temos, por exemplo, pontos geográficos que representam viagens de taxi. Cada ponto (1, 2, 3 e 4), possui seus atributos: preço da viagem, duração da viagem e a quantidade de passageiros; e sua localização geográfica. Na tabela 1, temos o exemplo

Ponto A	Ponto B	Relevância	Diversidade
1	2	0.8	0.9
1	3	0.25	0.2
1	4	0.9	0.45
2	3	0.2	1.0
2	4	1.0	0.48
3	4	0.2	0.3

Tabela 1: Exemplo de Índice de Relevância e Diversidade para pontos na Figura 3

do índice calculado diante dos pontos apresentados na Figura 3. Podemos perceber que os pontos mais similares entre si são os pontos 2 e 4, enquanto que 2 e 3 são os mais distantes, ou seja, possuem o maior valor de diversidade.

2.2.2 Preferências do Usuário

Para coletar as preferências do usuário, GeoGuide usa ambos feedback implícito e explícito. Feedback explícito é quando o usuário está analisando os atributos de um ponto, por exemplo a descrição de uma casa no Airbnb, e explicitamente pede para explorar pontos similares ao selecionado. Feedback implícito é coletado através da captura dos movimentos do mouse e métricas como “quanto tempo o usuário passou analisando o perfil de um ponto”.

2.2.3 Destacamento de Dados Espaciais

GeoGuide combina o índice preprocessado e o feedback coletado para destacar subconjuntos de dados espaciais de acordo com as preferências do analista. O processo de destacamento provou ser eficiente em termos de “quantos passos o analista leva até completar a tarefa de encontrar um ponto com um determinado perfil”. Usando GeoGuide, analistas foram capazes de completar a tarefa usando em média 10.7 passos, enquanto que usando Tableau, foram necessários 43 passos.

Neste trabalho, potencializamos GeoGuide à dois novos conceitos: *i.* regiões densas interessantes e *ii.* análise temporal das preferências do usuário.

3 Modelo Espaço-Temporal

Neste capítulo, entendemos o que são as chamadas Regiões Densas Interessantes e definimos o modelo espaço-temporal e sua aplicabilidade.

3.1 Regiões Densas Interessantes

Uma Região Densa Interessante (do inglês *Interesting Dense Region*, IDR) é uma região espacial com uma alta probabilidade de conter pontos de interesse do analista. IDR são coletadas e definidas durante o processamento do feedback do usuário. Diferentemente da literatura que predominantemente foca em interações explícitas, como clicar no botão, investigamos o feedback implícito.

Durante a exploração iterativa de dados espaciais, é comum o caso que o analista avalia algumas regiões de interesse, mas esquece de dar um feedback explícito sobre aquela região. O ato do usuário olhar para essa região pode ser capturado através do rastreamento dos movimentos oculares e, como Arapakis et al. (2014) comprova, esse método possui uma forte relação com a atenção do usuário.

Entretanto o rastreamento dos movimentos oculares fere várias questões de privacidade, assim sendo optamos pela alternativa de rastrear os deslocamentos do cursor do mouse. Arapakis, Lalmas e Valkanas (2014) argumenta que esse método possui uma forte relação com o engajamento do usuário. Intuitivamente, um ponto espacial recebe um feedback positivo se o cursor do mouse se desloca próximo a ele frequentemente.

O objetivo de descobrir IDRs é para obter preferências do analista que nunca foram expressadas explicitamente. Para tal, baseamos em dois conceitos:

- **Conceito 1:** uma região é mais interessante ao analista, se for densa. Por exemplo, o analista movimenta seu mouse naquela região várias vezes.
- **Conceito 2:** é possível que o analista movimente seu mouse em qualquer lugar do

Entrada: Tempo atual t_c , pontos dos movimentos do mouse \mathcal{M}

Saída: IDR's encontradas \mathcal{S}

```

1  $\mathcal{S} \leftarrow \emptyset$ 
2  $g \leftarrow \text{número de segmentos}$ 
3 para  $i \in [0, g]$  faça
4    $\mathcal{M}_i \leftarrow \{m = \langle x, y, t \rangle | (\frac{t_c}{g} \times i) \leq t \leq (\frac{t_c}{g} \times (i + 1))\}$ 
5    $\mathcal{C}_i \leftarrow \text{mine\_clusters}(\mathcal{M}_i)$ 
6    $\mathcal{O}_i \leftarrow \text{find\_polygons}(\mathcal{C}_i)$ 
7 para  $\mathcal{O}_i, \mathcal{O}_j$  onde  $i, j \in [0, g]$  e  $i \neq j$  faça  $\mathcal{S}.append(\text{intersect}(\mathcal{O}_i, \mathcal{O}_j))$ 
8 retorna  $\mathcal{S}$ 

```

Algoritmo 1: Descobrimento de IDR's

mapa. Isso não deve significar que todo lugar no mapa tem a mesma significância.

Com base nesses conceitos, o processo de descobrimento das IDR's pode ser descrito pelo Algoritmo 1. Adicionamos pontos para \mathcal{M} a cada $200ms$ para evitar pontos redundantes. De acordo com Conceito 1 e com o propósito de identificar comportamentos recorrentes, o algoritmo começa particionando \mathcal{M} em g segmentos consecutivos \mathcal{M}_i , até \mathcal{M}_j . O primeiro segmento começa no momento zero (quando a iteração iniciou) e o último termina em t_c , ou seja, no tempo atual que o algoritmo é executado. De acordo com o Conceito 2, encontramos os clusters em cada segmento de \mathcal{M} usando uma variação da abordagem DB-SCAN (ESTER et al., 1996). Por fim, encontramos e retornamos as interseções desses clusters como IDR's.

Para clusterizar os pontos em cada segmento de tempo (linha 5 do Algoritmo 1), usamos ST-DBSCAN, uma variação do DB-SCAN que leva em consideração atributos espaciais e temporais para clusterizar pontos com base na densidade (BIRANT; KUT, 2007). Para cada subconjunto de pontos dos movimentos do mouse \mathcal{M}_i , $i \in [0, g]$, ST-DBSCAN começa com um ponto aleatório $m_0 \in \mathcal{M}_i$ e coleta todos os pontos alcançáveis (de acordo com a densidade) de m_0 usando uma métrica de distância. Como os pontos dos movimentos do mouse ocorrem em um espaço bidimensional (ou seja, monitor do computador), optamos por usar a distância euclidiana como nossa métrica. Se m_0 for identificado como objeto *core*, uma cluster será gerado. Caso contrário, se m_0 é uma objeto de borda, nenhum ponto é alcançável de m_0 e o algoritmo buscará outro ponto aleatório em \mathcal{M}_i . Esse processo é repetido até que todos os pontos sejam processados.

Uma vez que os clusters são formados para cada subconjunto de \mathcal{M} , encontramos suas interseções para determinar as regiões recorrentes. Para obter as interseções, precisamos definir nitidamente as fronteiras dos clusters (linha 6). Portanto para cada cluster, nós encontramos seu respectivo polígono que engloba todos os pontos. Para isso, utilizamos

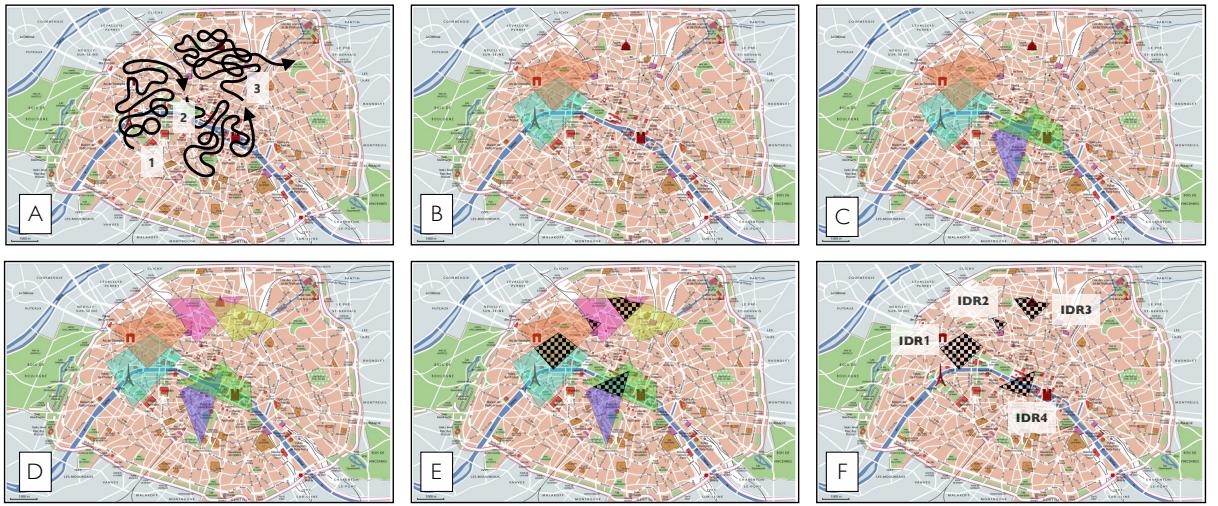


Figura 4: Processo de Descoberta de IDRs.

o algoritmo Quickhull, um método, com abordagem similar ao Quicksort, que calcula o invólucro convexo para um determinado conjunto de pontos num plano bidimensional (BARBER; DOBKIN; HUHDANPAA, 1996).

Esse algoritmo de descoberta de IDR é executado a cada t intervalo de tempo durante uma análise exploratória. Dessa forma, enquanto o analista está explorando o conjunto de dados, nosso sistema está coletando de maneira transparente suas preferências. Na Figura 4, vemos uma iteração de coleta e execução desse algoritmo onde $g = 3$, ou seja, o feedback é coletado em 3 segmentos diferentes de tempo. Os movimentos do mouse realizados pelo analista nesses 3 segmentos são demostrados na Figura 4.A. Para cada segmento, os pontos coletados dos movimentos do mouse são clusterizados e convertidos em polígonos convexos (Figura 4.B-D, respectivamente). Por fim, os polígonos são sobrepostos (Figura 4.E) e as intersecções encontradas são definidas como IDRs (IDR1-4 na Figura 4.F).

3.1.1 Perfil

Uma IDR descreve uma região espacial no conjunto de dados do analista coletada num determinado tempo t . Cada região representa as preferências do analista no momento t no contexto espacial, ou seja, *onde* o analista está interessado. Para entender, *em quê* o analista está interessado, em outras palavras, as preferências sobre o contexto de domínio, definimos o perfil de cada região.

O perfil é a sumarização dos atributos dos pontos contidos numa IDR. Cada ponto em um conjunto de dados ($p \in \mathcal{P}$) é descrito por suas coordenadas espaciais e uma série de atributos de domínio $dom(p)$. A Tabela 2 apresenta um exemplo dos atributos de um

id	15
latitude	48.88880
longitude	2.320465
name	Nice appartment in Batignolles
host_name	Daniele
neighbourhood	Batignolles-Monceau
room_type	Entire home/apt
price	65
minimun_nights	1
stars	4
number_of_reviews	25
reviews_per_month	1.74
availability_365	35
last_review	2015-11-14T09:07:02+00:00

Tabela 2: Exemplo de atributos de um ponto que representa uma estadia ponto no conjunto de dados do Airbnb.

Na Tabela 2, os atributos *latitude* e *longitude* são as coordenadas espaciais, enquanto que todos os outros atributos caracterizam os atributos de domínio. Dentre os atributos de domínio, podemos encontrar 4 tipos de dados:

1. Numéricos: são atributos que representam valores numéricos, como, por exemplo, os atributos *price*, *minimun_nights*, *number_of_reviews*, *reviews_per_month* e *availability_365*.
2. Textuais: são atributos que representam valores textuais, como, por exemplo, os atributos *name* e *host_name*.
3. Categóricos: são atributos que representam valores tanto numéricos, quanto textuais, mas que se repetem pouco em todo o conjunto. Por exemplo, o atributo *stars* possue apenas cinco valores em todo o conjunto. Os atributos *room_type* e *neighbourhood* também possuem um número finito de opções, por isso são classificados como categóricos.
4. Temporais: são atributos que representam data e/ou hora de um determinado evento, como, por exemplo, o atributo *last_review* que representa quando foi publicado a última avaliação.

Esses 4 tipos são comumente encontrados em conjuntos de dados espaciais, portanto o perfil da IDR precisa representar de maneira resumida esses atributos em relação aos pontos nela contidos. Assim sendo, tratamos cada tipo de acordo com sua especificidade:

Atributo	Total	Média	σ	Mínimo	Máximo
price	73	106,68	62,62	40	401
minimun_nights	73	2,36	1,60	1	7
number_of_reviews	73	17,86	27,85	0	120
reviews_per_month	55	1,64	1,53	0,08	7,27
calculated_host_listings_count	73	4,05	10,83	1	56
availability_365	73	179,64	154,48	0	365

Tabela 3: Exemplo de perfil dos atributos numéricos de uma IDR

Atributo: name	
Termo	Ocorrências
champs	29
elysées	18
studio	14
near	13
flat	11

Tabela 4: Exemplo parcial de perfil do atributo textual de uma IDR

Categoria	Quantidade
<neighbourhood, "Batignolles-Monceau">	73
<room_type, "Entire home/apt">	69
<room_type, "Private room">	4
<stars, 1>	6
<stars, 2>	7
<stars, 3>	29
<stars, 4>	20
<stars, 5>	11

Tabela 5: Exemplo de perfil dos atributos categóricos de uma IDR

1. Numéricos: os atributos números encontrados em uma IDR são summarizados usando funções estatísticas como média, desvio padrão, mínimo e máximo. Podemos ver um exemplo do resultado dessa summarização na Tabela 3.
2. Textuais: os atributos textos são tratados separadamente. Para cada atributo, é feito o levantamento dos termos mais usados. Essa abordagem é utilizado por Kumar e Kaur (2017) para encontrar os termos mais usados no contexto temporal. A Tabela 4 apresenta os 5 primeiros resultados para o atributo *name* em uma IDR.
3. Categóricos: os atributos categóricos são contabilizados num dicionário M onde as chaves são uma tupla com o nome do atributo e o valor da categoria. Quanto maior a quantidade de repetições de uma categoria, mais relevante no contexto de uma IDR. A Figura 5 mostra o resultado num conjunto de dados para os atributos *neighbourhood*, *room_type* e *stars*.
4. Temporais: os atributos temporais são normalizados para uma escala linear (a partir do ano novo para datas e a partir da meia noite para horas) a fim de ser contabilizado as ocorrências. Por exemplo, uma atributo com a data *2018-12-21* é normalizado para *2018, 12, 355*, isto é, numa linha temporal esse evento ocorreu no ano 2018, no 12º mês e no 355º dia. As horas são normalizados de maneira similar, a hora *08:33:22* é normalizado para *8, 513, 30802*, isto é, esse evento ocorreu na 8º hora, 513º minuto e 30802º segundo. Com os valores normalizadas, podemos encontrar os dias e os momentos do dia mais recorrentes para determinado evento na IDR.

3.2 Modelo Temporal

Cada IDR é coletado num momento t e seu perfil definido através dos atributos de domínio dos pontos contidos na região delimitada. Dessa forma, para cada momento t em qualquer espaço de tempo, sabemos as preferências do usuário tanto no contexto espacial quanto no contexto de domínio.

O perfil pode ser definido para cada IDR, mas também para um conjunto de IDRs, ou seja, é possível realizar a análise das preferências do analista em qualquer espaço de tempo (minutos, horas, dias, semanas etc).

Assim sendo é através das IDRs coletadas que podemos aplicar a análise temporal e evoluir o sistema de destaque de informação do GeoGuide.

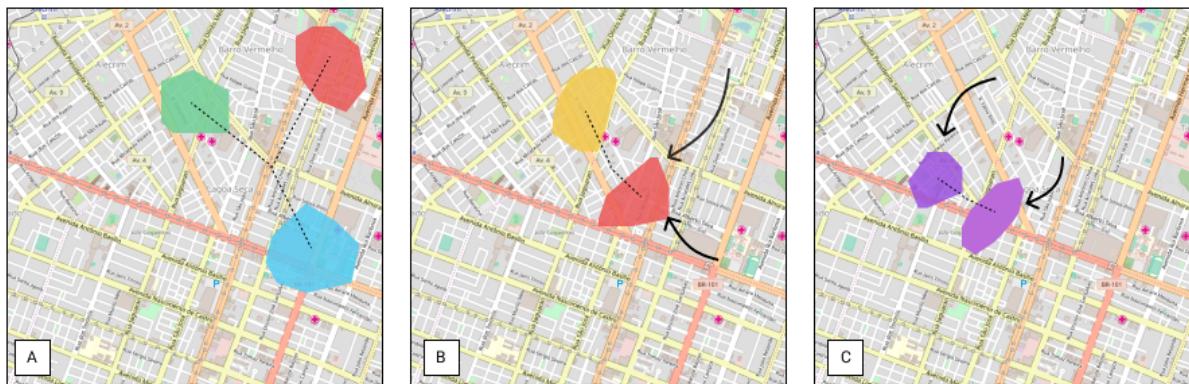


Figura 5: Evolução no contexto espacial.

3.3 Contexto Espacial

O contexto espacial se refere aos aspectos geográficos como, por exemplo, em que parte da cidade o analista se interessa.

3.4 Contexto de Domínio

O contexto de domínio se refere aos aspectos de domínio como, por exemplo, se o analista tem mais interesse em casas com varanda ou apartamentos.

4 Conclusão

TODO

4.1 Contribuições

TODO

4.2 Trabalhos futuros

TODO

Os dados coletados pelo nosso modelo podem ser utilizados para futuras análises envolvendo, por exemplo, grupos de analistas.

Referências

- ARAPAKIS, I. et al. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 65, n. 10, p. 1988–2005, 2014.
- ARAPAKIS, I.; LALMAS, M.; VALKANAS, G. Understanding within-content engagement through pattern analysis of mouse gestures. In: ACM. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. [S.l.], 2014. p. 1439–1448.
- 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>>.
- BARBER, C. B.; DOBKIN, D. P.; HUHDANPAA, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, ACM, New York, NY, USA, v. 22, n. 4, p. 469–483, dez. 1996. ISSN 0098-3500. Disponível em: <<http://doi.acm.org/10.1145/235815.235821>>.
- BIRANT, D.; KUT, A. St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 60, n. 1, p. 208–221, jan. 2007. ISSN 0169-023X. Disponível em: <<http://dx.doi.org/10.1016/j.datak.2006.01.013>>.
- 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>>.
- ESTER, M. et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <<http://dl.acm.org/citation.cfm?id=3001460.3001507>>.

- 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.; KŁOSOWSKI, 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>>.

PRADEEP, S.; KALLIMANI, J. S. A survey on various challenges and aspects in handling big data. In: *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*. IEEE, 2017. Disponível em: <<https://doi.org/10.1109/iceeccot.2017.8284606>>.

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.

ZHANG, J. et al. A survey of recent technologies and challenges in big data utilizations. In: *2015 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2015. Disponível em: <<https://doi.org/10.1109/ictc.2015.7354594>>.

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>>.