Encontrando os melhores locais a partir da popularidade de objetos de interesse na vizinhança espacial: uma proposta

Cláudio Moisés Valiense de Andrade[1], **João B. Rocha-Junior**[1]

[1]Universidade Estadual de Feira de Santana

claudiovaliense@gmail.com, joao@uefs.br

Abstract. *Spatial data is increasingly present in our daily lives. We use various applications that use this data, such as Google Maps and Uber. For example, a user wishing to know how many restaurants exist in their neighborhood may perform a spatial query, giving his current location (latitude and longitude) and a radius (e.g. 100m). This query return all restaurants inside the given region. This project proposes a new query type named Popularity based spatio-textual query that can select the best spatial objects taking into account the number of relevant spatio-textual objects, for a given set of query keywords, in its neighborhood. We present algorithms to process this query efficiently and evaluate the algorithms proposed in real datasets.*

1. Palavras-chave

Banco de Dados Espacial, Consulta Espaço-textual, Consulta preferencial.

2. Aluno

Cláudio Moisés Valiense de Andrade

3. Orientador

João B. Rocha-Junior

4. Nível

Mestrado Acadêmico

5. Nome do Programa de Pós-graduação e Universidade

Pós-Graduação em Computação Aplicada - Universidade Estadual de Feira de Santana

6. Ano/semestre de ingresso no programa

2017.1

7. Época esperada de conclusão

2018.2

8. Etapas já concluídas

Apresentação da proposta de dissertação do mestrado (14/12/2017)

9. Etapas futuras

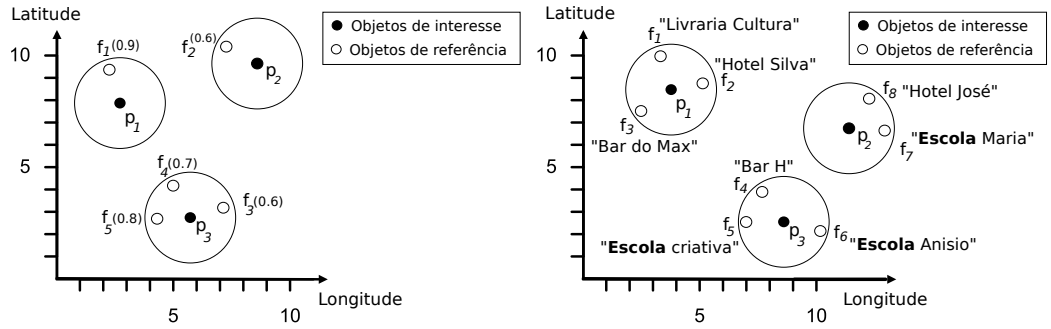
- Qualificação do mestrado (10/08/2018)
- Defesa da dissertação (10/12/2018)

1. Introdução

No nosso dia a dia, utilizamos diversas aplicações que manipulam dados, por exemplo, estamos dirigindo pela cidade e utilizamos o aplicativo *Waze*, este informa em qual parte do trajeto existem buracos ou sinalizações. O aplicativo está realizando consultas espaciais sobre os dados, gerando informações para atender a intenção de pesquisa do usuário.

Uma consulta espacial que vem sendo bastante estudada é a consulta espacial preferencial [Yiu et al. 2007]. Dado um conjunto de objetos espaciais de interesse (*locais*) e um conjunto de objetos espaciais de referência (*feature*), esta consulta retorna os k melhores objetos de interesse, levando-se em consideração o maior escore entre os objetos de referência dentro da região espacial de interesse fornecida pelo usuário. Nesta consulta, cada objeto de referência tem um escore que é definido por um provedor de classificação específico (e.g. ZAGAT[1], IFOOD[2]).

A Figura 1 exemplifica a consulta espacial preferencial. Assumindo que um usuário deseja ficar hospedado em um hotel que tenha o melhor bar considerando o entorno do hotel(eg. raio de 100m de distância). Analisando a Figura 1(a), dado o conjunto *p* de objetos de interesse (hotéis) e *f* objetos de referência (bar), o círculo delimita a região espacial de interesse (raio). Ao utilizar a consulta espacial preferencial para retornar os dois melhores hotéis (k=2), a consulta retorna os hotéis p_1 em primeiro e p_3 em segundo. O hotel p_1 é retornado em primeiro, porque no seu entorno tem um bar f_1 com escore 0.9, enquanto que o melhor bar no entorno de p_3 é f_5 com escore 0.8. Esta consulta considera apenas o escore do melhor objeto de referência dentro da região espacial de interesse.



(a) Consulta espacial preferencial.

(b) Consulta espaço-textual preferencial por popularidade.

Figura 1. Consultas espaciais preferenciais.

A consulta espacial preferencial tradicional tem dois problemas principais: 1) não leva em consideração a quantidade de objetos de referência na vizinhança espacial, mas apenas o escore do melhor objeto de referência. Por exemplo, um usuário esta interessado em ficar hospedado em um hotel que o seu entorno tenha a maior variedade de restaurantes; 2) não é possível selecionar os objetos de referência à partir da descrição textual dos objetos. A consulta espacial preferencial tradicional assume que o escore dos objetos são números estáticos. Entretanto, para a maioria das aplicações estes objetos estão associados a um texto descritivo. Neste caso, o escore de um objeto ou a sua relevância

[1]www.zagat.com

[2]www.ifood.com.br

deveria ser computada, levando-se em consideração a similaridade textual destes objetos com palavras-chave de busca.

A consulta proposta nesta pesquisa trabalha com dados espaço-textuais, denominada de Consulta Espaço-Textual Preferencial por Popularidade (CETPP). Diferente da consulta tradicional de [Yiu et al. 2007] que retorna o objeto de interesse analisando o objeto de referência com maior escore no seu entorno, a CETPP leva em consideração todos os objetos de referência que tem relevância textual maior que o mínimo pré-definido. A partir disso, o escore do objeto de interesse é baseado na contagem desses objetos de referência que ultrapassaram o limiar textual e que estejam abaixo da distância máxima informada. Esta pesquisa não tem como um dos objetivos, analisar a qualidade da resposta.

A Figura 1(b) é composta por objetos espaço-textuais. Para exemplificar o uso da consulta CETPP, assumimos que um cliente deseja comprar um apartamento que tenha muitas escolas na sua proximidade, ele pode então fornecer uma palavra-chave de interesse "escola" e indicar a região que considera próxima (ex. 1km). Ao realizar esta consulta na base de dados da Figura 1(b), ele tem como resposta o objeto p_3 com melhor escore, visto que p_3 possui 2 objetos na sua vizinhança espacial que satisfaz a palavra-chave de busca, enquanto que p_2 tem apenas 1 e p_1 não tem nenhum. Nesta consulta o escore de objeto de referência é o calculo de similaridade textual com as palavras-chave de busca.

2. Problema de pesquisa e caracterização da contribuição

Definição. Dado um conjunto de objetos de interesse P, onde cada objeto $p \in P$ possui uma coordenada espacial $p = (p.x, p.y)$; e um conjunto de objetos espaço-textuais de referência F, onde cada objeto $f \in F$ possui uma coordenada espacial $(f.x, f.y)$ e um texto $f.D, f = (f.x, f.y, f.D)$. A Consulta Preferencial por Popularidade Q tem 4 parâmetros, $Q = \{Q.D, Q.r, Q.k, Q.\sigma\}$, onde $Q.D$ é o conjunto de palavras-chave de interesse, $Q.r$ é o limiar que define o valor máximo da distância entre um objeto de interesse e de referência (raio), $Q.k$ é o número de resultados esperados e $Q.\sigma$ é o limiar que define o valor mínimo de similaridade textual que um objeto de referência pode ter para ser considerado textualmente relevante.

A consulta Q retorna o $Q.k$ objetos em P com os maiores escores. O escore de um objeto p, representado por $\tau(p)$, é a soma dos objetos de referência presentes na vizinhança espacial de interesse e que são textualmente relevantes para as palavras-chave de busca. A equação à seguir descreve essa especificação:

$$\tau(p) = \sum\{f \in F \mid dist(p, f) \leq Q.r : \theta(f.D, Q.D) > Q.\sigma\}$$

onde $\theta(f.D, Q.D)$ é a relevância textual (similaridade textual) entre o texto do objeto espaço-textual de referência $f.D$ e a palavras-chave de consulta $Q.D$. Nesta pesquisa, nós computamos a relevância textual como definida por [Rocha-Junior et al. 2011] e usamos a distância *Haversine* $dist(p, f)$ entre um objeto p e um objeto espaço-textual de referência f.

3. Trabalhos relacionados na área

Maximum Influence Optimal Influence Query [Du et al. 2005, Xia et al. 2005]. Esta consulta busca o ponto de máxima influência ao adicionar um novo objeto de interesse.

Este novo objeto de interesse precisa atender a maior quantidade de objetos de referência. Por exemplo, dado alguns locais de abertura de uma nova loja da *McDonald's*, um empresário deseja abrir uma franquia no local que atenda uma maior quantidade de clientes. O resultado desta consulta é este local. Uma das diferenças para o presente trabalho, é na utilização de palavras-chave para filtrar os objetos de referência.

Minimum Distance Optimal Location Query [Zhang et al. 2006]. Esta consulta busca o local que minimize a distância média de cada objeto de referência ao novo local de interesse. Por exemplo, um empresário deseja abrir uma franquia do *McDonald's* no local que diminua a distância média de todos os clientes até a loja *McDonald's* mais próxima. O presente trabalho utiliza de palavras-chave para filtrar os objetos de referência.

Top-k Spatial Keyword Preference Query [de Almeida and Rocha-Junior 2016]. Busca os top-k objetos de interesse de acordo com a relevância textual dos objetos de referência presentes na sua vizinhança espacial. Utiliza as palavras-chave para filtrar os objetos de referência. Por exemplo, um usuário esta interessado em alugar um apartamento próximo a um objeto relevante para as palavras-chave "escola" e "infantil", então é retornado os top-k apartamentos. A diferença para o presente trabalho, está no fato de que o escore dos objetos de interesse não são calculados a partir da contagem dos objetos de referência que atenderam ao critério espacial e textual.

4. Estado atual do trabalho

Os Algoritmos 1 e 2 processa a consulta CETPP retornando o mesmo conjunto de saída. No Algoritmo 1, é calculado o escore de cada objeto de interesse a partir dos objetos de referência que atenderam ao critério de espacial e textual. Dado o conjunto P de objetos de interesse, F o conjunto de objetos de referência, o conjunto $Q.D$ de palavras-chave, $Q.r$ que representa a distância máxima de um objeto de referência, $Q.k$ a quantidade de retorno de objetos de interesse e $Q.\sigma$ que representa o escore mínimo que um objeto de referência precisa ter para ser considerado. Na linha 1 é criada a MinHeap H para armazenar os elementos que tem o menor escore no topo. Linha 2 percorre o conjunto de objetos de interesse, com o objetivo que para cada objeto de interesse é realizado a contagem dos objetos de referência. Na linha 5 é calculado a distância , feito isso é verificado se a distância é menor ou igual que definida pelo usuário, na linha 6 é calculado o escore do objeto de referência em relação a similaridade textual com as palavras-chave *Q.D*, após calculado é testado se o escore do objeto é acima do limiar mínimo, caso seja, este objeto de referência entra na contagem. No cálculo da complexidade do Algoritmo 1, considere que n e m são as representações do conjunto dos objetos de interesse e referência. Este algoritmo tem complexidade $\Theta(n * m)$ por ter que percorrer todos os objetos de referência para cada objeto de interesse.

De forma similiar ao Algoritmo 1, o Algoritmo 2 utiliza o índice de dados espaciais *R-tree* para armazenar o conjunto dos objetos de referência. Na linha 5 é a realizado a consulta na *R-tree*, na qual é retornado todos os objetos de referência que esteja no limite da vizinhança espacial do objeto p, a partir disto é calculado o escore do objeto de interesse p. O restante do algoritmo funciona de forma análoga ao Algoritmo 2. A consulta na *R-tree* tem complexidade $\Theta(\log n)$ como mostrado por [Arge et al. 2008]. Como é preciso passar por cada objeto de interesse e realizar a consulta na *R-tree*, o algoritmo fica com complexidade $\Theta(n \log m)$.

Algoritmo 1: Popularity Spatial Keywords Preference Query (PSKPQ)

Input: P, F , $Q.D$, $Q.r$, $Q.k$, $Q.\sigma$
Output: MinHeap k objects

```
1  MinHeap H ← ∅;
2  forall p ∈ P do
3      count = 0;
4      forall f ∈ F do
5          if dist(p,f) ≤ Q.r then
6              if θ(f.D, Q.D) ≥ Q.σ then
7                  count++;
8      if count > 0 then
9          H.add(p, count);
10         if H.size() > Q.k then
11             H.remove();
12 return H;
```

Algoritmo 2: Popularity Spatial Keywords Preference Query R-tree (PSKPQR)

Input: P, F, $Q.D$, $Q.r$, $Q.k$, $Q.\sigma$
Output: MinHeap k objects

```
1  MinHeap H ← ∅;
2  R ← RTree(F);
3  forall p ∈ P do
4      count = 0;
5      foreach f ∈ R.search(p.x,p.y,Q.r) do
6          if θ(f.D, Q.D) ≥ Q.σ then
7              count++ ;
8      if count > 0 then
9          H.add(p, count);
10         if H.size() > Q.k then
11             H.remove();
12 return H;
```

Figura 2. Algoritmos para processar a CETPP.

5. Desenvolvimento necessário para a conclusão

Será preciso desenvolver novos algoritmos para processar a consulta de forma eficiente e comparar os resultados desta pesquisa com trabalhos similares. Testar em outras bases de dados disponíveis. Escrever a dissertação e artigos para serem avaliados por especialistas.

6. Avaliação dos resultados

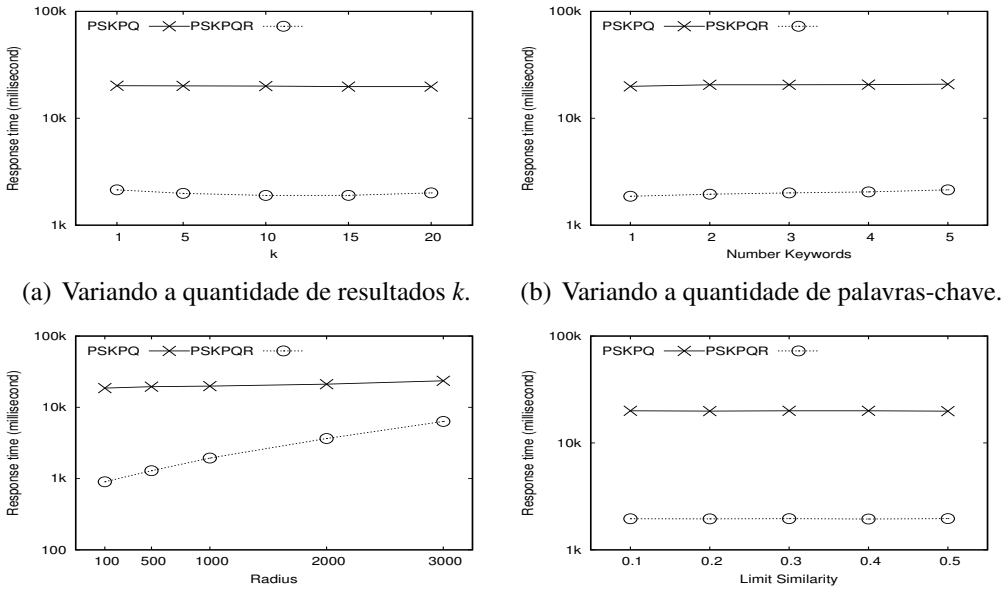
Na Tabela 1 é mostrado os parâmetros que serão variados nos experimentos.

Parâmetros	**Valores**
Número de resultados (k)	1, 5, **10**, 15, 20
Número de palavras-chave	1, 2, **3**, 4, 5
Valores do raio(metros)	100, 500, **1000**, 2000, 3000
Escore de similaridade textual	0.1, 0.2, **0.3**, 0.4, 0.5
Bases de dados	FSA, **SSA**, SP

Tabela 1. Parâmetros utilizados na consulta com os valores padrões destacados em negrito.

Neste experimento foi utilizado a base de dados *Open Street Map*, um sistema que fornece informações sobre localizações geográficas sobre locais existentes e ruas. Esta base representa os objetos de interesse como restaurantes, bares, hotéis, etc. Para representar os objetos de referência, foi utilizada a base do *Twitter*. Os objetos de referência

da base de dados do *Twitter* representa pessoas que enviaram mensagens através da rede social, esta contida nesta mensagem uma localização (latitude, longitude). A Figura 3 apresenta o processamento da consulta realizada por 2 algoritmos desenvolvidos na pesquisa. Nas Figura 3(a), 3(b) e 3(d) é possível observar que o algoritmo *PSKPQR* é cerca de 10 vezes mas rápido que o algoritmo básico que não utiliza índice. Na Figura 3(c), o algoritmo *PSKPQR* é sensível a variação do raio, na implementação da R-tree quanto menor o valor do raio, é retornado um conjunto menor de objetos de referência, com a tendência apresentada no gráfico, para um valor alto do raio, o algoritmo *PSKPQR* apresentará tempo de resposta maior que o *PSKPQ*.



(a) Variando a quantidade de resultados k.

(b) Variando a quantidade de palavras-chave.

(c) Variando o raio da consulta.

(d) Variando a similaridade.

Figura 3. Base de dados OSM e Twitter. Fonte: Próprio autor.

Referências

Arge, L., Berg, M. D., Haverkort, H., and Yi, K. (2008). The priority r-tree: A practically efficient and worst-case optimal r-tree. *TALG*, 4(1):9.

de Almeida, J. P. D. and Rocha-Junior, J. B. (2016). Top-k spatial keyword preference query. *Journal of Information and Data Management*, 6(3):162.

Du, Y., Zhang, D., and Xia, T. (2005). The optimal-location query. In *International Symposium on Spatial and Temporal Databases*, pages 163–180. Springer.

Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., and Nørvåg, K. (2011). Efficient processing of top-k spatial keyword queries. In *SSTD*, volume 6849 of *LNCS*. Springer.

Xia, T., Zhang, D., Kanoulas, E., and Du, Y. (2005). On computing top-t most influential spatial sites. *VLDB*, pages 946–957.

Yiu, M. L., Dai, X., Mamoulis, N., and Vaitis, M. (2007). Top-k spatial preference queries. In *Proceedings - ICDE*, pages 1076–1085. IEEE.

Zhang, D., Du, Y., Xia, T., and Tao, Y. (2006). Progressive computation of the min-dist optimal-location query. *VLDB*, pages 643–654.