Consulta Espacial Preferencial por Palavra Chave

João Paulo Dias de Almeida[1] **e João Batista da Rocha Júnior**[1]

[1]Pós-graduação em Computação Aplicada - Universidade Estadual de Feira de Santana (UEFS)
Feira de Santana - Bahia - Brasil

jpalmeida.uefs@gmail.com, joao@uefs.br

Abstract. *The Spatial Keyword Preference Query (Consulta Espacial Preferencial por Palavra Chave in Portuguese) is a new query type that aims to locate objects of interest, looking the spatial neighborhood of these objects. This query receives as parameter a set of objects of interest (e.g. hotels), a definition of neighborhood of interest (e.g. radius of 100m from the hotels) and a set of keywords of preference ("Mexican bar"); and returns the k best hotels ranked in terms of their vicinity with other spatial objects whose text is relevant for the query keywords. To the best of our knowledge, this query type has never been proposed before and it is an interesting tool to new spatial information systems. This paper presents a Master Thesis Project that aims to specify this new query type, presenting and evaluating new techniques to process it efficiently.*

Palavras-chave

objetos espaciais, consulta preferencial, consulta por palavra-chave

Aluno

João Paulo Dias de Almeida

Orientador

João Batista da Rocha Júnior

Nível

Mestrado

Ano/semestre de ingresso no programa

2013.2

Época esperada de conclusão

2015.1

Etapas já concluídas

Conclusão das disciplinas do 1° semestre (20/12/2013)

Etapas futuras

- Conclusão das disciplinas do 2° semestre (29/07/2014)
- Defesa da proposta de dissertação (28/07/2014)
- Qualificação (22/12/2014)
- Defesa da dissertação (11/05/2015)

1. Introdução

A geolocalização fornece comodidade e rapidez a todos que necessitem definir um trajeto para ir a um determinado local. Gorman (2008) identificou que a utilidade e velocidade do desenvolvimento de aplicações baseadas na geolocalização cresce exponencialmente com o volume de dados disponíveis. Devido a este mercado crescente, novas pesquisas estão sendo desenvolvidas com o objetivo de otimizar, ou desenvolver, novas técnicas de busca por objetos no espaço [Zhiming et al. 2012].

O objetivo desta pesquisa é desenvolver uma nova consulta, capaz de localizar pontos de interesse para o usuário. Para identificar a relevância de um ponto de interesse, a vizinhança espacial deste ponto é visitada à procura de outros objetos espaciais que satisfaçam as necessidades do usuário. Esta consulta é chamada de Consulta Espacial Preferencial por Palavra Chave (EPPC) e nunca foi proposta anteriormente na literatura.

Um usuário que deseja alugar um apartamento próximo a uma escola infantil (estabelecimento) pode utilizar a EPPC para encontrar os melhores apartamentos de acordo com a relevância textual entre as palavras-chave de busca e a descrição textual dos estabelecimentos na região de interesse. A Figura 1 representa um espaço que contém apartamentos (p) e estabelecimentos (f). Cada estabelecimento está associado a uma descrição textual. Por exemplo, o texto associado a $f3$ é "escola infantil". Nesta figura, a área de interesse é definida através de um raio em volta de cada apartamento. Sendo assim, o melhor apartamento é aquele que tem um objeto spatio-textual (estabelecimento) na sua vizinhança espacial, cujo texto é mais relevante para as palavras-chave de busca.

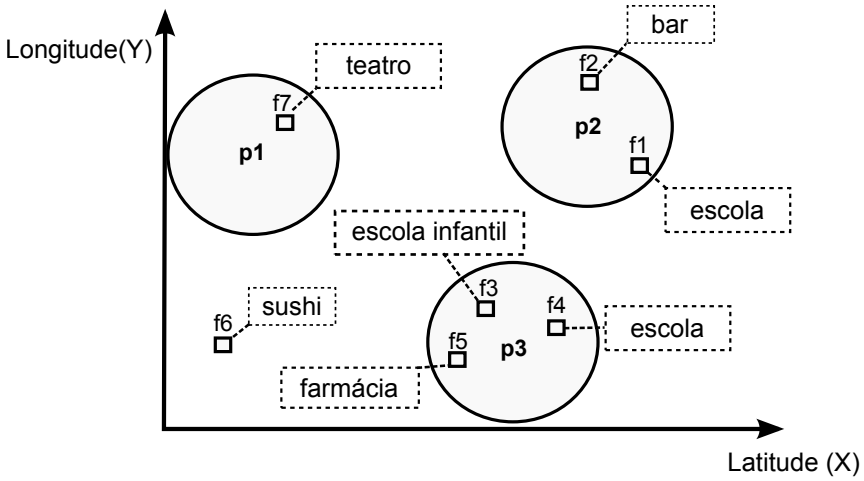

Figura 1. Área espacial contendo objetos de interesse (p) e estabelecimentos(f)

Por exemplo, sendo as palavras-chave de busca igual a "escola infantil", o escore do apartamento $p1$ é definido pelo cômputo da relevância textual entre a descrição de $f1$ ("escola") e as palavras-chave de busca "escola infantil". Sendo assim, $p2$ é mais relevante que $p1$, visto que na vizinhança de $p1$ não existe nenhum objeto textualmente relevante para as palavras-chave de busca. Neste exemplo, $p3$ é o objeto mais relevante, uma vez que o texto associado a $f3$ é mais textualmente relevante para consulta do que o texto associado a $f1$.

Este artigo é organizado como segue: a Seção 2 apresenta os trabalhos relacionados, destacando as principais diferenças entre o EPPC e as consultas já existentes na literatura. Na Seção 3 é feita a definição da consulta, e a Seção 4 descreve a situação atual do projeto. A Seção 5 apresenta as etapas necessárias para a conclusão do projeto, enquanto a Seção 6 apresenta os resultados obtidos. As conclusões são feitas na Seção 7.

2. Fundamentação teórica e trabalhos relacionados

Atualmente existem muitos trabalhos envolvendo consultas espaciais. Entre elas, as mais relevantes para a definição da EPPC são o *Spatial Keyword Querying* [Gao Cong 2012, Rocha-Junior et al. 2011] e o *Top-k Spatial Preference Queries* [Yiu et al. 2007, Rocha-Junior et al. 2010].

2.1. Top-k Spatial Preference Queries

Dado um conjunto D de objetos de interesse, uma *Top-k Spatial Preference Query* [Yiu et al. 2007] retorna os k objetos em D com os maiores escores. O escore de cada objeto de interesse é computado a partir do escore dos estabelecimentos presentes na sua vizinhança. O escore destes estabelecimentos é um valor fixo pré-definido, como por exemplo, a qualidade do objeto obtida em sistemas de qualificação como o http://www.zagat.com/.

A EPPC propõe uma consulta mais dinâmica, permitindo que o escore do estabelecimento não seja fixo, e sim computado dinamicamente a partir das palavras-chave de busca definidas no momento da consulta. Através desta consulta, o usuário é capaz de expressar que tipo de estabelecimento é relevante para as suas necessidades. A consulta avalia o quanto a descrição textual do estabelecimento é relevante para o conjunto de palavras-chave, e gera um escore representando esta relevância.

2.2. Spatial Keyword Queries

Uma *Spatial Keyword Query* [Gao Cong 2012] utiliza a localização do usuário e as palavras-chave, fornecidas por ele, como parâmetros. A consulta identifica objetos que são espacialmente próximos à localização do usuário, e textualmente relevantes às palavras-chave; retornando os k objetos próximos ao usuário e com maior relevância textual.

Para localizar objetos que sejam textualmente relevantes, a EPPC não leva em consideração a localização do usuário, e sim, um conjunto de objetos do interesse definidos pelo usuário. Assim, quando o usuário deseja procurar por estabelecimentos na vizinhança espacial dos seus objetos de interesse (e.g. hotéis), é verificada a relevância textual entre as palavras-chave de busca e a descrição textual de estabelecimentos próximos aos objetos de interesse.

3. Definição do Problema

Uma *Consulta Espacial Preferencial por Palavra Chave* (EPPC) Q é formada por um conjunto de palavras-chave e pela definição da vizinhança de interesse: $Q = (Q.s, Q.\tau, Q.k)$. Onde $Q.s$ representa as palavras-chave de busca, $Q.\tau$ o critério de vizinhança e $Q.k$ define o número de objetos de interesse que a consulta retorna ao usuário.

Dado um conjunto de objetos de interesse D, um conjunto de estabelecimentos F (*features*) e a *query* Q, a consulta retorna os k melhores objetos contidos em D que possuem maior escore.

O objeto p e o estabelecimento f são definidos como $p = (p.x, p.y)$ e $f = (f.t, f.x, f.y)$. Sendo x e y as coordenadas geográficas dos objetos p e f, enquanto $f.t$ se refere à descrição textual do *feature*. Esta descrição textual é comparada com as palavras-chave $Q.s$ informadas pelo usuário na *query* Q para computar o escore do objeto de interesse p. O escore do objeto p é definido através da relevância entre $Q.s$ e $f.t$, onde f é um *feature* presente na vizinhança espacial de interesse e cujo o texto $f.t$ é o mais relevante para a consulta $Q.s$.

Podendo a vizinhança ser definida por: *range* (rng), *nearest neighbor* (nn) ou *influence* (inf):

- O *range* escore $\tau^{rng}(p)$, dado um raio r,
$$\tau^{rng}(p) = max\{ W(Q.s, f.t) \mid f \in F, p \in D, d(p, f) <= r \}$$

- O *nearest neighbor* $\tau^{nn}(p)$,
$$\tau^{nn}(p) = max\{ W(Q.s, f.t) \mid f \in F, \forall v \in F : d(p, f) \leq d(p, v) \}$$

- Ou o *influence* escore $\tau^{inf}(p)$, considerando uma dada região r,
$$\tau^{inf}(p) = max\{ W(Q.s, f.t) \cdot 2^{-d(p,f)/r} \mid f \in F \}$$

Onde $W(Q.s, f.t)$ é a função que retorna a relevância textual entre o texto associado a *feature* f e o conjunto de palavras-chave $Q.s$. A distância entre um objeto p e um estabelecimento f é computada pela distância Euclidiana entre estes pontos ($d(p, f)$).

Quando o usuário opta por *range*, o escore do objeto de interesse é definido pelo *feature* mais relevante textualmente que está dentro da vizinhança especificada. Ao optar pelo *nearest neighbor*, o escore é definido pelo *feature* mais próximo do objeto de interesse que seja textualmente relevante. O *influence* define o escore do objeto de interesse a partir dos escores dados aos *features* na sua vizinhança. O escore de cada *feature* deve diminuir ou aumentar, de acordo com a distância entre o *feature* e o objeto de interesse. O maior escore existente na vizinhança é atribuído ao objeto de interesse.

4. Estado atual

Esta seção descreve o estado atual da pesquisa. Primeiramente, optamos por desenvolver um *Baseline* para processar a consulta EPPC utilizando apenas o *range* (critério de vizinhança). Com este *Baseline* é possível analisar o custo, no pior caso, para processar esta consulta. Posteriormente, iniciou-se o desenvolvimento de uma nova abordagem para processar esta consulta utilizando um índice spatio-textual.

4.1. *Baseline*

O *Baseline* computa o escore de cada objeto p, onde $p \in D$ e $f \in F$, retornando os k objetos com maior escore. Para computar o escore de cada objeto, o *Baseline* computa a distância espacial entre p e cada *feature* f existente na base de dados. Quando a distância $d(p, f)$ for menor do que $Q.r$, é computado a relevância textual $W(f.t, Q.s)$ entre f e Q.

O maior escore computado entre todos os *features* em F, onde $d(p, f) < Q.r$, é atribuído ao objeto p de interesse.

Esse processo se repete para cada p existente em D, mantendo os k-melhores objetos de interesse em uma *heap* H. Quando todo objeto $p \in D$ tenha sido visitado, é retornado os k-melhores objetos p armazenados na *heap* H.

4.2. Utilizando o índice

Um índice spatio-textual permite buscar pelos estabelecimentos f que atendem a vizinhança de interesse $Q.\tau$ e que sejam mais textualmente relevantes para as palavras-chave de busca $Q.s$ de forma eficiente. Desta forma, pode-se beneficiar deste índice para processar a consulta EPPC de forma mais eficiente.

Para cada p existente em D, é realizada uma busca no índice spatio-textual. A busca retorna o estabelecimento f existente na vizinhança de p cujo a descrição textual $f.t$ seja mais relevante para as palavras-chave de busca $Q.s$.

O escore de cada estabelecimento retornado pela busca no índice spatio-textual é atribuído ao escore do p objeto de interesse vizinho. O p objeto é inserido em uma *heap* onde são mantidos os k-melhores objetos de interesse. Este processo se repete até que a vizinhança espacial de todo objeto p existente em D seja visitada. Por fim, os k-melhores objetos de interesse mantidos na *heap* H são retornados como resposta da consulta.

Esta abordagem foi implementada utilizando o S2I [Rocha-Junior et al. 2011] como índice spatio-textual e tem se mostrado mais eficiente que o *Baseline*. Esta solução é similar ao *Simple Probing* proposto por Gao Cong (2012).

5. Próximos Passos

Inicialmente a consulta foi implementada utilizando apenas um critério de vizinhança (*range*). Portanto, pretende-se implementar essa consulta para os outros critérios de vizinhança: *nearest neighbor* e *influence*.

A solução baseada no índice S2I permite melhorar o desempenho da consulta. Entretanto, pretende-se investigar novos algoritmos que permitam processar essa consulta de forma diferente. Com isto, será possível analisar qual algoritmo de indexação apresenta o melhor desempenho.

Por fim, serão realizadas experimentações para avaliar a eficiência da consulta desenvolvida, como variar o tamanho do índice spatio-textual, variações no tamanho da base de dados ou a quantidade de palavras-chave.

6. Avaliação dos resultados

A avaliação dos resultados será realizada comparando os diversos algoritmos desenvolvidos para processar esta consulta em diferentes bases de dados. Os principais parâmetros utilizados na avaliação são: tempo de resposta, tamanho do índice, tamanho da base de dados e quantidades de palavras-chave.

6.1. Base de dados

Atualmente está sendo utilizada uma base de dados extraída do OpenStreetMap[1] (OSM) representando a cidade de Feira de Santana. No OSM, cada objeto espacial possui uma categoria (e.g. restaurante, bar ou hotel) e está a associado a um texto. A partir desta base, retira-se um conjunto de objetos de uma categoria para formar o conjunto de interesse D, utilizando-se os demais para compor o conjunto F.

Pretende-se ainda montar outras bases de dados com diferentes características como, por exemplo, o Wikipédia, onde os objetos espacias são associados a um texto mais representativo.

7. Conclusão

A consulta Espacial Preferencial por Palavra Chave (EPPC) é uma nova proposta para localizar objetos spatio-textuais. Esta pesquisa tem como finalidade a especificação da EPPC, e do desenvolvimento de técnicas capazes de processar a consulta de forma eficiente.

Através desta consulta é possível desenvolver novas ferramentas que podem ser utilizadas em Sistemas de Informação que auxiliem os usuários na tomada de decisões.

Referências

Gao Cong, Xin Cao, L. C. (2012). Spatial keyword querying. In *Conceptual Modeling*, pages 16–29. Springer.

Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., and Nørvåg, K. (2011). Efficient processing of top-k spatial keyword queries. In *Advances in Spatial and Temporal Databases*, pages 205–222. Springer.

Rocha-Junior, J. B., Vlachou, A., Doulkeridis, C., and Nørvåg, K. (2010). Efficient processing of top-k spatial preference queries. In *Proceedings of the VLDB Endowment*.

Yiu, M. L., Dai, X., Mamoulis, N., and Vaitis, M. (2007). Top-k spatial preference queries. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 1076–1085. IEEE.

Zhiming, C., Arefin, M. S., and Morimoto, Y. (2012). Skyline queries for spatial objects: A method for selecting spatial objects based on surrounding environments. In *Networking and Computing (ICNC), 2012 Third International Conference on*, pages 215–220. IEEE.

[1] http://www.openstreetmap.org/