PAPER

Evaluation of Top-k Spatial Keyword Preference Query

João P. D. ALMEIDA[†], **Frederico A DURÃO**[†], ***and*** **João B. ROCHA-JUNIOR**[††],

SUMMARY Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

key words: ***key words:*** *recommendation; ensemble; metadata; collaborative filtering*

1. Introduction

Avaliações são empregadas em Sistemas de Recuperação da Informação para aferir a performance ou qualidade do sistema. Como tal, avaliações são aceitas como uma necessidade critica para a ciência e tecnologia [1] pois a partir das avaliações é possível indicar quais são as fragilidades de um determinado sistema. Desta forma, este artigo apresenta uma avaliação qualitativa da Top-k Spatial Keyword Preference Query. Nesta avaliação qualitativa pretende-se mensurar a qualidade da resposta obtida pela Top-k Spatial Keyword Preference Query.

2. Related Work

3. The Top-k Spatial Keyword Preference Query

A consulta *Espacial Preferencial por Palavra-chave* (EPPC) Q é formada por um conjunto de palavras-chave $Q.D$, pela definição da vizinhança espacial de interesse $Q.\psi$, e pela quantidade $Q.k$ de objetos de interesse que se deseja obter como resposta: $Q = (Q.D, Q.\psi, Q.k)$.

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 Q retorna os k objetos de interesse contidos em P que possuem os maiores escores. O escore $\tau^{Q.\psi}(p)$ de um objeto de interesse $p \in P$ é a maior relevância textual (similaridade textual) entre os objetos espaço-textuais de referência f que atendem ao critério de vizinhança espacial $Q.\psi$..

Manuscript received January 1, 2015.
Manuscript revised January 1, 2015.
[†]The author is with the
DOI: 10.1587/trans.E0.??.1

O usuário pode optar entre três formas possíveis de especificar o critério de vizinhança espacial $Q.\psi$: seleção espacial ($Q.\psi = rng$), vizinho mais próximo ($Q.\psi = nn$) ou influência ($Q.\psi = inf$) [**?**].

- Seleção espacial $\tau^{rng}(p)$, dado um raio r:

$$\tau^{rng}(p) = max\{\theta(f.D, Q.D) \mid f \in F : dist(p, f) <= r\}$$

- Vizinho mais próximo $\tau^{nn}(p)$:

$$\tau^{nn}(p) = max\{\theta(f.D, Q.D) \mid f \in F, \theta(f.D, Q.D) > 0, \forall v \in F : dist(p, f) \leq dist(p, v)\}$$

- Influência $\tau^{inf}(p)$, dado um raio r:

$$\tau^{inf}(p) = max\{\theta(f.D, Q.D) \cdot 2^{-dist(p,f)/r} \mid f \in F\}$$

onde $\theta(f.D, Q.D)$ é a função cosseno que retorna a relevância textual (similaridade textual) entre o texto do objeto de referência $f.D$ e o conjunto de palavras-chave $Q.D$ [**?**], [**?**], enquanto $dist(p, f)$ é a distância Euclidiana entre um objeto p e um estabelecimento f.

Quando o usuário opta pela *seleção espacial* ($Q.\psi = rng$), o escore de um objeto de interesse $p \in P$ é definido pela maior relevância textual $\theta(f.D, Q.D)$ entre os objetos $f \in F$ que atendem ao critério de vizinhança espacial, ($dist(p, f) \leq r$). Ao optar por *vizinho mais próximo* ($Q.\psi = nn$), o escore é definido pelo objeto de referência mais próximo de p que seja textualmente relevante $\theta(f.D, Q.D) > 0$. Caso existam vários objetos de referência f com a mesma proximidade espacial do objeto de interesse p, o escore de p é a maior relevância textual $\theta(f.D, Q.D)$ entre os objetos f que possuem a mesma menor distância. O critério de vizinhança *influência* ($Q.\psi = inf$) define o escore do objeto de interesse p levando em consideração a relevância textual entre $f.D$ e $Q.D$ e a distância entre p e f, quanto maior a distância menor o escore (influência).

4. Experimental Methodology

É possível avaliar a qualidade da resposta da consulta EPPC utilizando julgamentos, uma base contendo objetos espaciais de interesse e uma base contendo objetos espaço-textuais de referência f. Os julgamentos determinam se um objeto de referência f é textualmente relevante para o conjunto de palavras-chave da consulta $Q.D$.

Não foi encontrada nenhuma base de dados que possuísse julgamentos e que possuísse as características necessárias para avaliar a consulta EPPC. Para tal, a base deve possuir objetos espaço-textuais (objetos formados por texto e coordenadas geográficas) e julgamentos que indiquem para quais palavras-chave um determinado objeto espaço-textual é relevante.

Julgamentos são usualmente encontrados em bases de dados direcionadas para sistemas que realizam Categorização Textual. Porém, estas bases de dados são formadas principalmente por objetos textuais, como a Reuters-21578[†], 20NewsGroups[††] ou WebKB[†††].

Desta forma, foi necessário construir a nossa própria base de dados. Para isto, utilizamos a base de dados Reuters-21578, Distribution 1.0, que contém documentos textuais categorizados e é amplamente utilizada para avaliar sistemas de recuperação textual [4].

Cada documento textual existente na Reuters-21578 foi categorizado por um indexador humano, responsável por indicar a que categoria[†] um determinado documento textual pertence.

Cada categoria está contida em um conjunto de categorias que agrega categorias semanticamente semelhantes. Os nomes destes conjuntos são: exchanges, orgs, people, places e topics. Sendo assim, o conjunto "places" contém todas as categorias relacionadas ao tópico "lugar". Por exemplo, uma categoria denominada "australia" está contida no conjunto "places".

A associação entre categoria e documento textual é feita quando o conteúdo textual do documento é relacionado semanticamente a uma determinada categoria. Desta forma, um documento que pertence às categorias "usa" e "uruguay" possui um texto semanticamente relevante para o termo "usa" e também para o termo "uruguay". A seguir é apresentado um exemplo de documento que pertence as categorias "usa" e "uruguay".

Exemplo. Cake sales were registered at 785 to 995 dlrs for March/April, 785 dlrs for May, 753 dlrs for Aug and 0.39 times New York Dec for Oct/Dec. Buyers were the U.S., Argentina, Uruguay and convertible currency areas.

Entretanto, para que seja possível executar a consulta EPPC utilizando a base Reuters-21578 é necessário que os documentos textuais também possuam coordenadas espaciais (latitude e longitude). Então foi atribuída uma coordenada espacial para cada documento textual existente na Reuters-21578.

Assim, uma das categorias é selecionada ("usa" ou "uruguay") e todo documento que pertence a esta categoria é posicionado na vizinhança espacial de pelo menos um objeto de interesse. Ao executar a consulta EPPC, o termo que representa a categoria é utilizado como palavra-chave. Ou seja, se uma categoria denominada "grain" é selecionada, o termo "grain" é utilizado como palavra-chave na consulta EPPC. Assim garantimos a existência de documentos espaço-textuais de referência que são relevantes para a palavra-chave e que também estão na vizinhança espacial dos objetos de interesse.

Utilizando similaridade cosseno, é realizado o calculo de relevância textual entre os documentos espaço-textuais que pertencem à categoria selecionada e a palavra-chave fornecida à consulta EPPC. A partir dos valores obtidos, é construído um *rank* destes documentos espaço-textuais em ordem decrescente. Este rank é denominado de "judgment rank".

O resultado obtido pela consulta EPPC é comparado com o judgement rank. O valor do escore do objeto de interesse mais relevante retornado pela consulta EPPC deve ser o mesmo valor que está no topo do judgment rank.

Exemplo: A Figura 1 apresenta uma área espacial contendo objetos de interesse p e objetos de referência f. Os objetos de referência são documentos textuais da Reuters-21578, enquanto os objetos de interesse são representados por objetos espaciais obtidos do OpenStreetMap[†].

Dada uma consulta Q, cujo o conjunto de palavras-chave seja $Q.D$ = "$ship$", a consulta EPPC identifica qual objeto de interesse possui, em sua vizinhança espacial, objetos de referência relevantes para o conjunto de palavras-chave $Q.D$. Os objetos f_1 e f_2 estão na vizinhança espacial do objeto de interesse e foram categorizados pela Reuters-21578 como pertencentes à categoria "ship", portanto estes objetos são relevantes textualmente para o conjunto de palavras-chave $Q.D$. Observe que mesmo o objeto f_1 não possuindo este termo explicitamente em sua descrição textual, ele é considerado relevante para o termo "ship" pois o conteúdo do texto é relacionado ao termo. Os demais objetos (f_3, f_4, f_5) não estão na vizinhança espacial de nenhum objeto de interesse, por isto a descrição textual destes objetos não é analisada pela consulta EPPC.

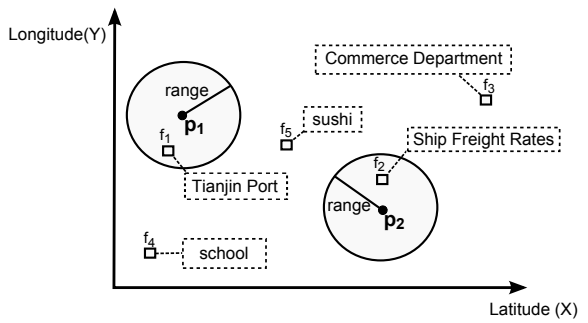

Fig. 1 Área espacial contendo objetos de interesse e objetos de referência

Utilizando a similaridade cosseno, calcula-se a relevância textual de cada documento textual e sua categoria (ex: a relevância entre a descrição textual do objeto f_1 e o termo "ship"). Em seguida é construído um rank composto por

[†]Disponível em: http://goo.gl/NXNxnq

[††]Disponível em: http://qwone.com/ jason/20Newsgroups/

[†††]Disponível em: www.cs.cmu.edu/ TextLearning/datasets.html

[†]Os detalhes de como esta categorização foi realizada são apresentados por Lewis [3]

[†]Disponível em: https://mapzen.com/data/metro-extracts/

estes objetos e eles são ordenados decrescentemente pelo valor obtido durante o calculo da relevância textual. Denominamos este rank de "JudgmentRank(t)", onde t corresponde aos termos existentes no conjunto de palavras-chave $Q.D$.

Espera-se que ao executar a consulta EPPC com um conjunto de palavras-chave $Q.D = t$, obtenha-se uma resposta igual ao rank "JudgmentRank(t)". Assim, garante-se que a resposta da consulta é composta apenas por documentos textuais relevantes.

Então, para a avaliar a

Os documentos textuais existentes na Reuters-21578 não possuem coordenadas espaciais. Sendo assim, acrescentamos um par de coordenadas (latitude e longitude) a cada documento textual. Um grupo destes documentos textuais é posicionado na vizinhança espacial de um objeto de interesse. Assim garantimos quais documentos são espacialmente relevantes para um objeto de interesse.

Para garantir a relevância textual de um documento para uma query Q, verifica-se se a categoria deste documento (fornecida na Reuters-21578) coincide com o conjunto de palavras-chave da query $Q.D$.

5. Evaluation

6. Conclusion

References

[1] S. Tefko. Evaluation of evaluation in information retrieval. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995.

[2] A. Töscher, M. Jahrer, and R. M. Bell. The bigchaos solution to the netflix grand prize. *Netflix prize documentation*, 2009.

[3] D. Lewis. Representation and learning in information retrieval. *University of Massachusetts*, 1992.

[4] D. Lewis. Reuters-21578 Test Collection. *Reuters-21578 Documentation*, http://goo.gl/NXNxnq, 2016.