Ferramenta para recuperação de informação utilizando indexação espacial e textual

Mairon Q. Castro[1], **Clodoveu A. Davis Jr.**[2]

[1,2] Departamento de Ciencia da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{mairon.castro,clodoveu}@dcc.ufmg.br

Abstract. *Search engines usually focus on keyword-based search, and thus require searches related to places to be resolved using place names among the keywords, with mixed results. This paper describes the structure and development of an information retrieval engine that allows the user to search by terms and by geographic limits. Searching combines a term index and a geographic index, allowing results to reflect a combination of both interests. We demonstrate the effectiveness of the approach using a dataset of 1.7 million georeferenced tweets, collected during the 2014 World Cup.*

Resumo. *Máquinas de busca em geral focam a busca baseada em palavras-chave, e portanto exigem que buscas relacionadas a lugares sejam realizadas com base em nomes geográficos, obtendo resultados de qualidade variável. Este artigo apresenta a estrutura e implementação de um mecanismo de recuperação de informação capaz de combinar termos e limites geográficos na entrada, obtendo resultados que refletem a combinação das estratégias. A eficiência do enfoque proposto é demonstrada em buscas sobre um conjunto de 1.7 milhão de tweets georreferenciados, coletados ao longo da Copa do Mundo de 2014.*

1. Introdução

Máquinas de busca são recursos bastante comuns e presentes na vida cotidiana das pessoas. A partir de um conjunto de palavras-chave, máquinas de busca localizam e classificam fontes de conteúdo online referentes ao que lhes parece ser a intenção de busca do usuário.

No entanto, quando existem referências a lugares entre as palavras-chave, muitas vezes a máquina de busca não consegue capturar a intenção de restringir a busca ou relacionar os resultados obtidos a determinado local. Seria interessante, portanto, que o usuário pudesse expressar sua intenção de busca através de referências geoespaciais mais diretas, apontando um lugar em um mapa ou delimitando uma região de busca, além de informar palavras-chave que pudessem definir tematicamente o interesse de busca.

Para isso, seria interessante contar tanto com uma interface de consulta quanto com mecanismos de indexação que tratassem tanto da parte geoespacial quanto da parte temática de uma consulta. Além disso, a classificação ou ranqueamento dos resultados poderia combinar aspectos dos componentes geoespacial e temático, expressando uma noção de preferência do usuário nos resultados obtidos.

Este artigo apresenta um método de indexação espacial e textual de dados coletados na Web, e descreve um protótipo de ferramenta que permite fazer buscas híbridas, espaciais-textuais. São exploradas diferentes modalidades de consulta e de apresentação de resultados, tendo em vista um potencial uso das propostas aqui apresentadas na indexação espacial-textual de conjuntos de documentos de diferentes naturezas, tais como documentos científicos ou notícias cotidianas.

2. Trabalhos Relacionados

O projeto SPIRIT (Spatially-Aware Information Retrieval on the Internet) [Jones et al. 2002] explora conceitos de busca espacial e textual. Ao longo de sua existência, o projeto SPIRIT desenvolveu uma

máquina de busca completa para localizar documentos e conjuntos de dados relacionados a lugares ou regiões referenciadas em uma busca. Aspectos do projeto abordaram desde a identificação de referências a lugares nos documentos e reconhecimento de referências espaciais entre as palavras-chave da busca, além da criação de índices e de funções de ranqueamento.

Outro projeto relacionado ao tema deste artigo é o Somewherenear [Banahan et al. 2000]. Trata-se de uma ferramenta de busca geográfica que permite que usuários localizem itens de interesse em proximidade a outros itens, com base na distância. O intuito principal é servir como fonte de informação para viajantes a lazer ou a negócios, em busca de lugares para visitar, acomodação, alimentação e outros serviços.

3. Componentes da ferramenta

3.1. Índice Espacial

Um índice geográfico ou espacial é responsável por recuperar informação sobre objetos espaciais, cuja forma geométrica e localização apresentam mais de uma dimensão. Assim, o índice deve ser capaz de, dado um par de coordenadas, ou a delimitação geográfica de uma região, retornar objetos que estão contidos neste limite. Sistemas de bancos de dados tradicionais usam índices unidimensionais (ou seja, baseados em um atributo ou chave), como árvore B e hash, para resolver consultas de forma eficiente. Índices convencionais não são suficientes para dados geográficos, pois consultas espaciais exigem a recuperação eficiente de dados considerando mais de uma dimensão, e também considerando proximidade, topologia, dimensões e outras características dos objetos geográficos, tipicamente codificados segundo pontos, linhas e polígonos.

Uma das estruturas de indexação espacial mais utilizadas em bancos de dados geográficos é a R-tree[Guttman 1984]. A R-tree utiliza o retângulo envolvente mínimo como representação simplificada da geometria dos objetos, e indexa retângulos. Cada nó da árvore representa um retângulo que contém todos os retângulos que descendem dele. Em um mesmo nível da R-Tree é possível que os retângulos de nós irmãos apresentem superposições, o que gera a necessidade de o algoritmo de indexação lidar com o problema de agregação e subdivisão de retângulos, buscando aumentar a eficiência do índice. Numerosas propostas de variação da política de agregação e subdivisão de retângulos foram apresentadas na literatura, gerando variações da R-Tree. Mesmo assim, a versão original é a mais usualmente empregada pelos gerenciadores de bancos de dados geográficos.

3.2. Índice Textual

No contexto de uma máquina de busca tradicional, que opera por palavras chave, é criado um índice ou arquivo invertido, em que palavras e expressões são relacionadas aos documentos onde foram encontradas. Quando o usuário fornece uma lista de palavras para sua busca, a máquina de busca recupera as listas de referências a documentos presentes no índice e obtém a interseção entre essas listas, ou seja, documentos relacionados a todas as palavras da busca. O índice invertido é, assim, responsável por organizar dados que possibilitem que a busca seja feita a partir de palavras, sem que haja acesso direto à coleção de documentos. Uma busca sequencial, sem qualquer tratamento dos dados inviabilizaria a priorização do conjunto resposta.

A utilização do arquivo invertido aumenta a eficiência de pesquisa em várias ordens de magnitude, característica importante para aplicações que utilizam grandes volumes de documentos constituídos de texto. O custo para se conseguir essa eficiência é a necessidade de armazenar uma estrutura de dados que pode ocupar tanto espaço quanto o texto original, dependendo da quantidade de informação armazenada no índice [Manning et al. 2008a].

Existem várias estratégias para buscar a redução do custo de manutenção do índice. *Stopwords* são palavras muito comuns em uma linguagem, como artigos e preposições. O objetivo é não indexar *stopwords*, pois não trazem muita informação sobre um documento. Um termo indexável é uma palavra que não seja *stopword* presente em um documento. Como a finalidade é recuperar informações de qualquer documento, um indexador deve percorrer a coleção, recuperar e armazenar todos os seus termos indexáveis, bem como informações sobre eles. O conjunto de todos os termos indexáveis únicos de uma coleção é chamado de *vocabulário*.

Com o vocabulário em mãos, é possível obter todos os pares termo-documento. Ordenando esses pares por termo e depois por documento, pode-se organizar todas as informações sobre um termo de forma contígua no dispositivo de armazenamento, otimizando a busca por este termo. Este conjunto ordenado de pares é o arquivo invertido de uma coleção. É de se notar que a ordenação deve ser realizada por alguma técnica de ordenação em memória secundária, devido ao grande tamanho do índice.

Portanto, a construção do vocabulário de uma coleção se baseia em percorrer todos os documentos e analisar palavra por palavra. Para cada documento, deve-se inserir sua chave na coleção. Após a identificação dos termos indexáveis, o indexador percorre todos e verifica se está no vocabulário ou não e insere caso não esteja. Se já estiver, apenas a frequência é atualizada. O mesmo é feito para o índice, verificando se o termo atual já possui entrada no índice para este documento.

O tamanho do índice cresce linearmente com o tamanho da coleção, fato que impossibilita armazenamento em memória primária de grandes coleções. Ao analisar a estrutura de cada entrada do índice, que é na forma da tupla (termo, documento, frequência), pode-se observar que após a leitura completa de um documento, entradas no índice relativas a este documento não serão mais atualizadas, pois consideramos que os documentos são únicos.

Existem estratégias para lidar com índices dinâmicos, ou seja, índices capazes de se adaptarem a mudanças em documentos, permitindo a reindexação dos mesmos[Manning et al. 2008a]. Este tipo de índice é útil para páginas que são atualizadas com frequência, como por exemplo páginas iniciais de portais de notícias. O índice dinâmico não é abordado neste trabalho.

3.3. Ranqueamento

Com o índice pronto, para as pesquisas terem resultados relevantes, funções de ranqueamento devem ser aplicadas, ou seja, é preciso calcular uma pontuação numérica para a associação entre consultas e documentos. No caso deste trabalho, isso precisa ser feito tanto para o texto, quanto para a representação geográfica associada ao documento.

3.3.1. Ranqueamento textual

Na busca textual, o primeiro passo é definir um modelo que dê pesos para pares termo-documento, ou seja, que descreva a importância de um termo para um documento. Com o modelo definido, é possível modelar as pesquisas e os documentos como vetores, nos quais cada posição é dada pelo valor do peso de cada termo único, tanto para a coleção quanto para o documento. O modelo $tf - idf_{t,d}$(*Term Frequency - Inverse Document Frequency*, onde t é um termo e d um documento) para atribuição de pesos é muito popular em recuperação de informação [Manning et al. 2008b].

O modelo tf se baseia no fato de que o peso de um termo é proporcional à sua frequência em um documento, ou seja, quanto mais frequente um termo é em um documento, maior é o seu peso. Isso se fundamenta na observação de que termos de alta frequência são importantes para descrever documentos.

Já $idf_{t,d}$ é importante porque, se um termo é muito frequente na coleção, e está em boa parte dos documentos, a probabilidade de ser um termo específico é baixa. Para isso, o modelo estabelece uma punição a termos muito frequentes, diminuindo o peso deles. Isto é, quanto menos frequente na coleção um termo da consulta for, maior deve ser o seu peso. Por exemplo, a consulta 'refrigerante de guaraná' deve dar um maior peso ao termo 'guaraná', porque é uma palavra que é mais específica do que 'refrigerante' no vocabulário português.

O $tf-idf_{t,d}$ é um modelo que alia os dois pontos, no intuito de chegar a uma forma coerente de pesos. O cálculo é dado pela seguinte expressão:

$$tf - idf_{t,d} \; = \; tf_{t,d} \times log(\frac{N}{df_t}) \tag{1}$$

Onde $tf_{t,d}$ é a frequência do termo no documento, N é o tamanho da coleção e df_t é a quantidade de documentos em que o termo aparece. Representando as consultas e os documentos em um modelo

de espaço vetorial [Manning et al. 2008b], onde os vetores são os pesos $tf_{t,d} \times idf_t$ entre a consulta e os documentos, a função de ranqueamento é dada pela similaridade de cosseno entre estes vetores.

$$sim(q,d) \ = \ v(\vec{q}).v(\vec{d}) \ = \ \frac{V(\vec{q}).V(\vec{d})}{\|V(\vec{q})\|.\|V(\vec{d})\|} \tag{2}$$

O ranqueamento é calculado através da ordenação dos documentos por ordem decrescente de pontuação, pois quanto maior o cosseno entre dois vetores, mais próximo os seus unitários estão. Isto é, mais próxima um documento está de uma consulta.

3.3.2. Ranqueamento geográfico

No caso da estratégia de ranqueamento para a busca por delimitação geográfica, não há um método que seja unanimidade, já que diversos fatores podem estar envolvidos [Kumar 2011]. Os limites geográficos da base de dados podem variar muito, o objetivo com a busca também. Geralmente as estratégias consideram medidas de similaridade espacial, como sobreposição, forma, contorno, etc.

Uma possibilidade para ranqueamento geográfico é o uso de alguma função de distância. Por exemplo, pode-se usar uma ordenação por proximidade a algum ponto de referência citado na consulta, ou a distância ao centro de uma região indicada para a consulta.

A interface da ferramenta desenvolvida possibilita ao usuário fazer uma delimitação geográfica através do desenho de um retângulo no mapa. Como os dados indexados estão representados apenas como pontos, a alternativa de ranqueamento implementada foi através da distância euclidiana do documento ao centro do retângulo delimitado na consulta, de forma que quanto mais perto do centro, maior é a relevância do documento. Assim sendo, se a consulta q for o retângulo delimitado por $(lng1, lng2, lat1, lat2)$ e o ponto de um documento d for (lng, lat), temos:

$$lngCenter = \frac{(lng1 + lng2)}{2} \tag{3}$$

$$latCenter = \frac{(lat1 + lat2)}{2} \tag{4}$$

$$sim(q,d) = \sqrt{(lng - lngCenter)^2 - (lat - latCenter)^2} \tag{5}$$

4. Ferramenta e coleção utilizada

A Figura 1 apresenta a estrutura da ferramenta implementada. O usuário apresenta termos de busca e/ou delimita uma região geográfica de seu interesse, utilizando um retângulo sobre um mapa. São realizadas consultas aos dois índices (textual e geográfico), e os resultados são combinados e ordenados, de acordo com a estratégia de ranqueamento descrita. Ao final, os resultados são apresentados sobre o mapa, e podem ser consultados individualmente.

A ferramenta desenvolvida[1] utiliza o gerenciador de bancos de dados PostgreSQL para armazenar o vocabulário, e emprega a extensão geográfica PostGIS para gerenciar objetos espaciais e geográficos e indexá-los, bem como para realizar operações com estes objetos. A estrutura de indexação usada pelo PostGIS é baseada na R-tree, já citada.

Já o arquivo invertido é armazenado no disco como um arquivo ordenado por termo e depois por documento, e contém a frequência do termo em cada documento. A ordenação é feita através do algoritmo de intercalação [Greene 1991].

A coleção de documentos utilizada para este trabalho, de modo a testar a ferramenta implementada, foi um conjunto de 1.715.167 tweets coletados ao longo da Copa do Mundo de 2014. Cada tweet está associado a uma posição (latitude e longitude). A coleta foi realizada por pesquisadores

[1] http://greenwich.lbd.dcc.ufmg.br/termgeo/

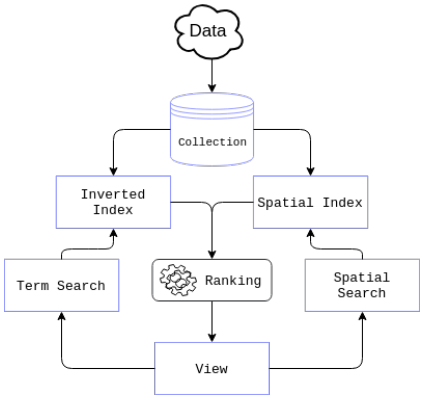

Figura 1. Estrutura do projeto

do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais. O conjunto original de tweets continha cerca de 50 vezes mais elementos, mas aqueles sem geolocalização foram descartados: apenas cerca de 2% dos tweets coletados estavam associados a um par de coordenadas.

```
Index I = {}
Collection C = {}
Vocabulary V = {}
for each document d do
    read document;
    insert d coordinates and key in C
    for each valid term t do
        if t is not in V then
            insert with frequency 1
        else
            update frequency
        end
        if tuple(t, d) is not in I then
            insert with frequency 1
        else
            update frequency
        end
    end
    if memory is full then
        save I, C, V on disk
        I = {}
    end
end
order I by (t, d)
```

Algorithm 1: Construção dos índices

O algoritmo 1 resume a construção dos índices da ferramenta. As estratégias de ranqueamento já citadas são utilizadas para cada consulta.

A Figura 2 apresenta imagens da interface da ferramenta durante a realização das diferentes buscas. Ao passar o cursor do mouse sobre os documentos do conjunto resposta, a ferramenta evidencia no mapa, o local exato do documento em foco. É possível também dar zoom e acessar cada um dos Tweets, clicando no documento.

5. Conclusões e trabalhos futuros

Analisando a ferramenta e seus resultados, é interessante notar que uma simples consulta por termos destaca justamente as regiões mais desenvolvidas do país no mapa. Deve-se ao fato da facilidade de

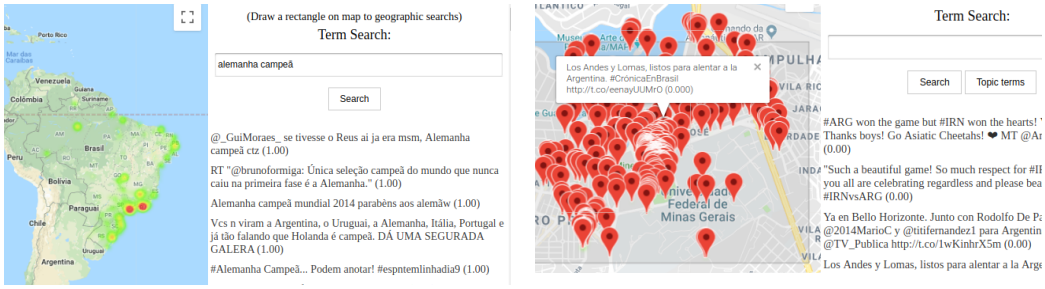

Figura 2. À esquerda, busca por termos - Pesquisa por "alemanha campeã". À direita, busca por delimitação geográfica no entorno do estádio Mineirão, em Belo Horizonte. Esse tipo de busca ativa o botão "Topic terms", que possibilita a visualização dos principais termos na região delimitada.

acesso à internet, e consequentemente, um maior uso de redes sociais como o Twitter. Também é possível notar que busca por delimitação geográfica destaca termos relacionados à Copa do Mundo, o que é esperado dada a delimitação feita e o período de coleta dos dados. Uma ferramenta desse tipo pode possibilitar diversas análises interessantes, como por exemplo popularidade de um político ou um time de futebol, ou até mesmo identificar focos de doenças, analisando a localização de mensagens contento de citações a elas. É possível também analisar o que está sendo falado em determinado local e o porquê. É de se ressaltar que a ferramenta é genérica e pode funcionar para qualquer conjunto de dados, desde que geolocalizados.

Tomando o $idf_{t,d}$ como exemplo, é possível também desenvolver uma estratégia de ranqueamento espacial semelhante, que leve em consideração a especificidade de uma determinada região, assim como o $idf_{t,d}$ leva à especificidade de um termo. Isto é, se uma determinada região possui poucos documentos, uma busca por delimitação geográfica que inclua esta e outras regiões com mais documentos deverá dar um maior peso para esta região. Isso pode ser interessante para destacar documentos de áreas desfavorecidas.

Outra proposta de trabalho seria utilizar a ferramenta para análises de buscas, no intuito de tirar conclusões sobre determinado assunto e/ou região.

Referências

Banahan, M., Fisher, D., Greenwood, A., Riach, J., and Willis, L. (2000). Somewherenear is the uk's leading geographic search engine. [Online; accessed 29-August-2018].

Greene, W. A. (1991). k-way merging and k-ary sorts. In *[Proceedings] 1991 Symposium on Applied Computing*, pages 197–.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *SIGMOD Rec.*, 14(2):47–57.

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., and Weibel, R. (2002). Spatial information retrieval and geographical ontologies an overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 387–388, New York, NY, USA. ACM.

Kumar, C. (2011). Relevance and ranking in geographic information retrieval. In *Proceedings of the Fourth BCS-IRSG Conference on Future Directions in Information Access*, FDIA'11, pages 2–7, Swindon, UK. BCS Learning & Development Ltd.

Manning, C. D., Raghavan, P., and Schütze, H. (2008a). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Manning, C. D., Raghavan, P., and Schütze, H. (2008b). *Scoring, term weighting, and the vector space model*, page 100–123. Cambridge University Press.