Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Avaliação Qualitativa da Consulta Espaço-Textual Top-*k*

Luiz Felipe Naziazeno Neto

Feira de Santana

2017

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Luiz Felipe Naziazeno-Neto

Avaliação Qualitativa da Consulta Espaço-Textual Top-*k*

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: João B. Rocha-Junior

Coorientador: Rodrigo Tripodi Calumby

Feira de Santana

2017

Ficha Catalográfica - Biblioteca Central Julieta Carteado

N245a Naziazeno Neto, Luiz Felipe
Avaliação qualitativa da Consulta Espaço- Textual Top-k / Luiz Felipe Naziazeno Neto. - 2017.
56 f.: il.

Orientador: João B. Rocha Júnior.
Coorientador: Rodrigo Tripodi Calumby.

Dissertação (mestrado) - Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2017.

1. Sistemas de recuperação da informação. 2. Consulta Espaço-Textual. I. Rocha Júnior, João B., orient. II. Calumby, Rodrigo Tripodi, coorient. III. Universidade Estadual de Feira de Santana. IV. Título.

CDU: 025.4.03

Luiz Felipe Naziazeno Neto

Avaliação Qualitativa da Consulta Espaço-Textual Top-k

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 04 de agosto de 2017

BANCA EXAMINADORA

Dr. João Batista da Rocha Junior (Orientador)
Universidade Estadual de Feira de Santana

Dra. Tatiane Nogueira Rios
Universidade Federal da Bahia

Dr. Matheus Giovanni Pires
Universidade Estadual de Feira de Santana

Abstract

It is increasingly common to find large databases with textual and spatial (latitude and longitude) information. Therefore, there is great interest in new strategies to efficiently retrieve relevant information from these datasets. One solution, which has attracted the attention of the research community, is the use of the Top-k Spatial-Textual Query. This query returns the best k spatio-textual documents considering the spatial distance between the query location and the documents and the textual relevance between the query keywords and the text associated with the documents. This work, however, studies the Top-k Spatial-Textual Query from a qualitative perspective, proposing a methodology for creating test collections using metrics to evaluate the quality of the results of the two existing ranking functions, from a textual and spatial perspective. The generated collections have a great diversity of spatial-textual data, which is vital for a qualitative study. The gained knowledge in this research provides an important source for those interested in evaluating information retrievals systems with spatio-textual characteristics.

Keywords: Qualitative Study, Spatio-Textual Query, Precision

Resumo

É cada vez mais comum encontrar grandes bancos de dados com informações textual e espacial (latitude e longitude). Portanto, existe um grande interesse por novas estratégias para recuperar eficientemente informações relevantes a partir destes conjuntos de dados. Uma solução, que tem atraído a atenção da comunidade científica, é a utilização da Consulta Espaço-Textual Top-k. Esta consulta retorna os k melhores documentos espaço-textuais considerando a distância espacial entre o local da consulta e os documentos e a relevância textual entre as palavras-chave da consulta e o texto associado aos documentos. As pesquisas anteriores concentram-se nos aspectos de performance dessa consulta, como a redução do tempo de resposta e I/O. Este trabalho, no entanto, estuda a Consulta Espaço-Textual Top-k a partir de uma perspectiva qualitativa, propondo uma metodologia para criacão de coleções de teste espaço-textuais, utilizando métricas para avaliar a qualidade dos resultados das duas funções de ranqueamento existente, sob a perspectiva textual e espacial. As coleções geradas proporcionaram uma grande diversidade de dados espaço-textuais, o que é vital para uma análise qualitativa. O conhecimento adquirido, nesta pesquisa, proporciona uma fonte importante aos interessados em avaliar sistemas de recuperação com características espaço-textuais.

Palavras-chave: Estudo Qualitativo, Consulta Espaço-Textual, Precisão

Prefácio

Esta dissertação de mestrado foi submetida a Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvido dentro do Programa de Pós-Graduacão em Computação Aplicada (PGCA) tendo como orientador o Dr. **João B. Rocha-Junior** e co-orientador o Dr. **Rodrigo Tripodi Calumby**.

Agradecimentos

Agradecer... Talvez este seja um dos maiores desafios deste trabalho! Como é importante reconhecer o apoio das pessoas para se alcançar qualquer que seja o objetivo.

Ao meu orientador, Professor Dr. João B. Rocha-Junior, exemplo de profissional dedicado e competente. Verdadeiro exemplo de batalha e superação, dia a dia, levando o nome da Universidade Estadual de Feira de Santana a lugares de destaque. Obrigado, porque alem dos ensinamentos cient´ıficos, o senhor demonstrou que acreditando nas pessoas comprometidas é possível realizar sonhos.

Ao meu co-orientador, Professor Dr. Rodrigo Tripodi Calumby, que alem de ter sido primordial para a conclusão dessa dissertação, é uma pessoal humilde com todos os seus alunos e colegas da nossa estimada Universidade Estadual de Feira de Santana.

Ao meu pai, José Felipe Naziazeno Neto, e minha mãe, Edilena Maria Menezes Naziazeno, por todo esforço que fizeram para me dar educação de qualidade. Obrigado, principalmente a minha mãe, que conviveu com todos os problemas ao longo desse mestrado.

À minha mulher Professora Msc. Maria Cecília Castelo Branco de Santana, que não tenho palavras para agradecer o empenho. Dedicou ao máximo nesse mestrado, prejudicando muitas vezes o seu doutorado para me ajudar.

Aos meus filhos Luana, João Vitor e João Marcelo, que são a razão do meu viver. Amor eterno.

Sumário

Lista de Publicações

- Uma proposta de estudo qualitativo da consulta espaco-textual top-k. Luiz F. Naziazeno-Neto e João B. Rocha-Junior. *In III Workshop de Trabalhos de Pós-Graduação da Erbase (WPOS)*, Maceió, 2016.
- Avaliação Qualitativa da Consulta Espaco-Textual Top-k. Luiz F. Naziazeno-Neto, João B. Rocha-Junior, Rodrigo Tripodi Calumby. *In XIII Simpósio Brasileiro de Sistemas de Informação (SBSI), 2017*.

Lista de Tabelas

Lista de Figuras

Lista de Abrebiações

Abreviação	**Descrição**
EQA	Equação relacionada à [Cong et al. 2009]
EQB	Equação relacionada à [Rocha-Junior e Nørvåg 2012]
ASS	*Average Spatial Similarity*

Capítulo 1

Introdução

"É muito melhor lançar-se em busca de conquistas grandiosas, mesmo expondo-se ao fracasso, do que alinhar-se com os pobres de espírito, que nem gozam muito, nem sofrem muito, porque vivem numa penumbra cinzenta onde não conhece nem vitória nem derrota."

– Theodore Roosevelt

É cada vez mais comum encontrar grandes bases de dados com informação textual e espacial (latitude e longitude). Isto explica o interesse por novas técnicas para recuperar informações relevantes destas bases de dados. Uma forma de extrair estas informações é através de consultas espaço-textuais. As consultas espaço-textuais estão auxiliando diversas aplicações, tais como, *Google Maps*[1], na qual pontos de interesse podem ser encontrados; *Foursquare*[2], na qual documentos georreferenciados com recomendações de lugares de interesse (Ex: bares e restaurantes) podem ser recuperados; *Twitter* [3], na qual *tweets* podem ser retornados [Chen et al. 2013].

Uma consulta espaço-textual que vem atraindo muita atenção da comunidade científica é a Consulta Espaço-Textual Top-k [Cong et al. 2009]. Esta consulta retorna os k melhores documentos ordenados por uma pontuação, que é calculada levando-se em consideração dois aspectos: a distância entre a localização do documento e a localização da consulta; e a similaridade textual entre o texto do documento e as palavras-chave da consulta.

[1]https://maps.google.com/
[2]https://foursquare.com/
[3]https://twitter.com/

A Figura 1.1 ilustra o funcionamento da Consulta Espaço-Textual Top-k. A consulta requer: o valor de k, que é o número de documentos retornados pela consulta, a localização da consulta q, representado na figura pelas coordenadas (x_q, y_q), e as palavras-chave da consulta, que para o exemplo da figura são representadas pelas palavras "italiana" e "chinesa". Para retornar os melhores documentos para o usuário, a consulta pontua cada um, representados na figura por p_1, p_2, p_3, p_4, p_5 e p_6, de acordo com a similaridade textual e a distância em relação ao local da consulta q. No que se refere à similaridade textual, quanto mais semelhante textualmente for um documento em relação às palavras-chave da consulta, maior será sua similaridade textual; no que se refere à proximidade espacial, quanto mais próximo o documento estiver ao local da consulta, maior será sua pontuação espacial. Os documentos da Figura 1.1 são representações de estabelecimentos próximos à consulta q.

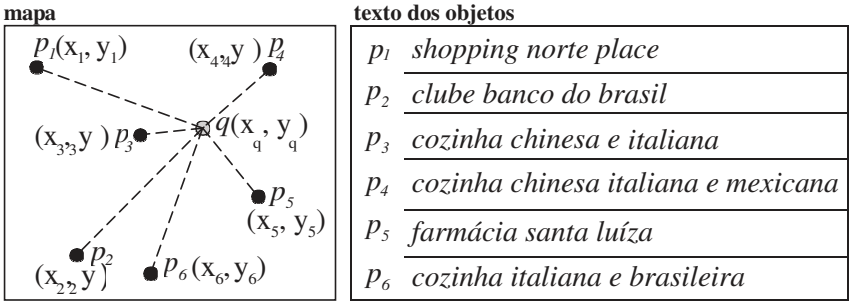


Figura 1.1: Execução da Consulta Espaço-Textual Top-k.

Supondo que os valores utilizados, pelo usuário deste exemplo, foram: $k = 2$ (ou seja, o usuário deseja que a consulta retorne os dois melhores estabelecimentos), palavras-chave = "*italiana*", "*chinesa*" e as coordenadas (x_q, y_q) da consulta. Assim, fazendo uma análise em todos os documentos da figura, os documentos que possuem as palavras-chave da consulta são os documentos p_3, o p_4 e p_6. Neste cenário, os dois documentos com maior similaridade textual e proximidade espacial são os documentos p_3 e o p_4, pois esses documentos possuem as duas palavras-chave da consulta ("italiana" e "chinesa"), além de estarem mais próximos ao local da consulta.

O resultado do exemplo acima foi produzido por uma função de ranqueamento [Salton e McGill 1986]. Atualmente, a Consulta Espaço-Textual Top-k possui duas funções de ranqueamento, [Cong et al. 2009] e [Rocha-Junior e Nørvåg 2012]. Elas têm como objetivo pontuar cada documento espaço-textual de acordo com a distância do documento em relação ao local da consulta (quanto mais perto um documento, maior é a sua pontuação) e a similaridade textual desse documento em relação às palavras-chave da consulta (quanto maior essa similaridade, maior é a sua pontuação), com o intuito de ranquear os documentos com as maiores pontuações [Rocha-Junior 2012, Cao et al. 2012, Chen et al. 2013].

Ainda sobre o exemplo anterior, os estabelecimentos representados pelos documentos p_3 e p_4 tiveram as melhores pontuacões, sugerindo que esses são os estabelecimentos

mais relevantes para o usuário. No entanto, o fato das palavras-chave da consulta estarem em um determinado documento ou esse documento ser mais próximo ao local da consulta, não necessariamente atende a uma necessidade de informação do usuário. A relevância de um resultado está associada à necessidade de informação, ela não está associada à consulta e é uma questão pessoal. Assim, um documento só deve ser considerado relevante se e somente se suprir a essa necessidade [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013]. Logo, os estabelecimentos p_3 e p_4 podem ser os mais relevantes para um determinado usuário, mas, esses documentos podem não ser relevantes para um outro usuário.

As pesquisas atualmente voltadas para a Consulta Espaço-Textual Top-k preocupam-se com o desempenho do processamento da consulta, como o seu tempo de resposta ou o espaço de memória utilizado [Rocha-Junior 2012, Cao et al. 2012, Chen et al. 2013]. No entanto, na perspectiva apresentada, esta pesquisa se pro- põe avaliar a relevância dos resultados obtidos pela Consulta Espaço-Textual Top-k utilizando as duas funções de ranqueamento existentes para esse tipo de consulta, avaliando a qualidade desses resultados em relação ao atendimento da necessidade de informação do usuário.

Para avaliar um resultado de uma consulta, quanto à relevância e, consequentemente, verificar se este resultado atende a uma necessidade de informação, existem duas abordagens: a centrada no comportamento do usuário diante de um sistema de recuperação e a centrada no sistema. As duas abordagens avaliam se o sistema de recuperação avaliado atende à necessidade de informação do usuário.

As abordagens centradas no sistema são as mais tradicionais e dominantes na literatura e são chamadas de abordagens *Cranfield* [Baeza-Yates e Ribeiro-Neto 2013]. Dispensa-se a influência de aspectos humanos durante a avaliação, pois utiliza-se coleções de teste preestabelecidas. Os experimentos são facilmente escaláveis e reproduzíveis, o que facilita uma comparacão quantitativa entre distintas implementações. As coleções de testes geradas nesta abordagem são padronizadas e facilitam a comparação entre funções de ranqueamento distintas.

As abordagens centradas no comportamento do usuário diante de um sistema de recuperação de informação focam em descobrir como as preferências dos usuários podem ser afetadas pelas caracter´ısticas da interface de um sistema e pela sua facilidade de uso [Wilkinson e Wu 2004], além de avaliar se a consulta atende a uma necessidade de informação. Além disso, levam em consideração a acessibilidade, desempenho do usuário, usabilidade, dentre outros aspectos. Esses métodos são derivados dos conceitos de IHC (Interação Humano-Computador).

Apesar de oferecerem um tratamento mais completo, os métodos centrados no usuário são experimentos difíceis de reproduzir, de projeto complexo e caros [Göker e Myrhaug 2008]. Além disso, as abordagens centradas no usuário podem ser muito dependentes da usabilidade da interface do sistema e da experiência do usuário com aplicações de recuperação de informação.

Devido as características descritas sobre as duas abordagens e levando em consideração o custo de implementação, este trabalho utiliza a abordagem centrada no sistema para avaliar a Consulta Espaço-Textual Top-k.

Nesta pesquisa, pretende-se responder as seguintes questões:

- Como avaliar o resultado da Consulta Espaço-Textual Top-k quanto à relevância?
- Como avaliar o resultado da Consulta Espaço-Textual Top-k quanto à proximidade espacial?
- Qual a qualidade dos resultados utilizando as métricas selecionadas?
- Qual função de ranqueamento que proporciona melhores resultados qualitativos?

Dada as questões de pesquisa, as principais contribuições são:

1. Especificação e implementação de um modelo de criação de coleções de teste espaço-textuais a partir de coleções tradicionais;
2. Uma metodologia para avaliacão da Consulta Espaço-Textual Top- k utilizando a abordagem centrada no sistema;
3. Definição de uma métrica para avaliar a distância dos documentos relevantes retornados em relação à localização da consulta.

Os próximos capítulos estão organizados da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica relacionada ao tema. O Capítulo 3 descreve a metodologia de avaliação e o Capítulo 4 apresenta os resultados. Por fim, o Capítulo 5 apresenta as considerações finais desta dissertação.

Capítulo 2 Fundamentação Teórica

"Quanto mais aumenta nosso conhecimento, mais evidente fica nossa ignorância."

– John F. Kennedy

Este Capítulo trata da revisão bibliográfica utilizada nesta pesquisa, consistindo em estudos sobre a área de Recuperação de Informacão, abordando, principalmente, aspectos que dão suporte para a avaliacão qualitativa da consulta. Além disso, é feito um estudo sobre consultas espaciais por palavra-chave e índices espaciais, que dão a base para o entendimento da Consulta Espaço-Textual Top-*k*.

2.1 Recuperação de Informação

O termo Recuperação de Informação (RI) foi criado por [MOOEM 1951], que definiu da seguinte forma: "*...Recuperação de Informação é o nome do processo onde um possível usuário de informação pode converter a sua necessidade de informação em uma lista de citações de documentos armazenados que contenham informações úteis a ele*".

A área de Recuperação de Informação (RI) tem como foco principal o usuário e suas necessidades para, principalmente, oferecer de maneira eficaz, o acesso à informação. Recuperar informação é encontrar material (comumente documentos) de um ambiente não estruturado (normalmente texto) dentro de grandes coleções, que satisfaça uma necessidade de informação [Manning et al. 2008].

Para um Sistema de Recuperação de Informações, quando um usuário informa uma consulta, inicia-se um processo de Recuperação de Informacão. As consultas são representações formais das necessidades de informacão de um usuário. Em um Sistema

de Recuperação de Informação uma consulta não é associada a um único documento em uma coleção, ao contrário, diversos documentos são recuperados através de uma consulta, selecionando-se os documentos que se apresentam como mais relevantes comparando a consulta com as representações dos documentos previamente armazenados [Manning et al. 2008].

A ideia de um sistema de RI é recuperar mais documentos relevantes e, consequentemente, menos documentos irrelevantes. Além disso, uma questão importante é que a relevância é um julgamento pessoal, que deriva de um problema a ser resolvido e do seu contexto. Por exemplo, à medida que novas informações vão sendo disponibilizadas, a relevância pode mudar [Baeza-Yates e Ribeiro-Neto 2013]. Ou seja, tempo é uma propriedade importante relacionada à relevância. Além disso, o local de interesse tem um impacto significativo na relevância (exemplo, um restaurante mais próximo pode ser a resposta mais relevante).

O usuário pode representar a sua necessidade de informação através de uma consulta. Essa necessidade de informação em geral não é especificada em uma linguagem natural e sim através de palavras-chave. Após a concepção da consulta, o Sistema de Recuperação de Informação tenta localizar documentos que possam ser relevantes para o usuário, com o objetivo de atender a uma necessidade de informação do usuário [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013].

Os Sistemas de Recuperação de Informação possuem uma função que compara a consulta fornecida pelo usuário com os textos armazenados no repositório. Esta função deve retornar o grau de relevância dos documentos em relação às consultas. Os documentos identificados com maior grau de relevância são mostrados primeiro para o usuário [Baeza-Yates e Ribeiro-Neto 2013].

Na maioria das vezes, o texto de um documento não é armazenado por inteiro em um Sistema de Recuperação de Informacão. Para cada texto são criadas estruturas de dados, como o arquivo invertido que será apresentado na próxima seção, com o objetivo de acelerar o seu processo de recuperação [Salton e McGill 1986, Zobel e Moffat 2006, Yan et al. 2009, Baeza-Yates e Ribeiro-Neto 2013]. Os textos que devem ser indexados são submetidos a um processo de filtragem de termos relevantes, denominada extração de atributos, que são utilizados para caracterizar os documentos armazenados [Manning et al. 2008].

2.1.1 Arquivo Invertido

No que se refere às informações textuais, é comum indexá-las através de um índice chamado Arquivo Invertido (*Inverted File*). Esse índice é constituído de uma lista ordenada (ou ´ındice) de palavras-chave (ou atributos), na qual cada palavra-chave tem uma lista de ponteiros para os documentos que as contêm [Zobel e Moffat 2006, Baeza-Yates e Ribeiro-Neto 2013]. O Arquivo Invertido possui dois componentes: lista invertida (*posting list*) e vocabulário (dicionário de termos). Um termo t existente na coleção tem uma lista invertida correspondente. Cada lista invertida possui

um identificador do documento (d_{id}) acompanhado da frequência $f_{t,D}$ com que o termo t aparece neste documento. O vocabulário armazena dois atributos: f_t, que é a quantidade dos documentos contendo t, e um ponteiro para o início da lista invertida correspondente [Zobel e Moffat 2006, Baeza-Yates e Ribeiro-Neto 2013].

termo t	*ft*	Lista invertida de *t*
maçã	1	(d_6, 2)
casa	2	(d_2, 2) (d_3, 1)
fruta	1	(d_6, 1)
banheiro	1	(d_4, 1)
porta	1	(d_2, 1)
lagoa	1	(d_3, 1)
dia	2	(d^2, 1) (d_3, 1)
noite	5	(d_1, 1) (d_2, 2) (d_3, 1) (d_5, 1) (d_6, 2)
frio	3	(d_1, 1) (d_3, 1) (d_5, 1)
praia	3	(d_1, 1) (d_4, 1) (d_5, 1)

Figura 2.1: Exemplo de ´ındice invertido [Zobel e Moffat 2006].

A Figura 2.1 representa um exemplo de um ´ındice invertido de um determinado banco de dados. A entrada de cada termo t é composta da frequência f_t e uma lista de pares, cada par composto de um documento identificado por d_i e a frequência $f_{d,t}$ do termo no documento.

2.1.2 Modelos de Recuperação de Informação

Um modelo de recuperação de informação é uma abordagem para resolver o problema de relevância de documentos. Existem diversos modelos de recuperação de informação para ordenar documentos de uma coleção e os mais conhecidos são: modelo vetorial, modelo probabilístico e o modelo estatístico de linguagem [Salton e McGill 1986, Baeza-Yates e Ribeiro-Neto 2013].

A atuação de um modelo de recuperacão de informação é basicamente sobre uma coleção de documentos e um conjunto de consultas. Além disso, o modelo possui uma função de ranqueamento que ordena os documentos recuperados, de maneira a organizar os documentos mais relevantes no topo. Uma caracterização formal dos modelos foi proposta por Baeza (2011), como mostra as definições abaixo[Baeza-Yates e Ribeiro-Neto 2011]:

Definição 1. Um modelo de recuperacão de informacão é definido por uma quádrupla $[D, Q, F, r(q_i, d_j)]$ onde

- D é um conjunto composto por representacões para os documentos em uma coleção;

- Q é um conjunto formado por representacões (consultas) para uma necessidade de informação do usuário;
- F é um arcabouço para modelagem de representacões de documentos, consultas e seus relacionamentos;
- $r(q_i, d_j)$ é uma função de ordenacão que associa um número real a uma consulta $q_i \in Q$ e uma representação de documento $d_j \in D$ para ordenar os documentos de acordo com a consulta.

Definição 2. Seja D uma coleção de documentos e $d_j \in D$ um documento pertencente à coleção D. Seja Q um conjunto de consultas e $q_i \in Q$ uma consulta para qual existe um conjunto de documentos relevantes pertencentes à coleção D. Define-se função de ranqueamento como sendo a função r tal que $r(q_i, d_j) : Q \times D \rightarrow \mathrm{R}$, representando a medida de similaridade entre uma consulta q e um documento d_j.

As funções de ranqueamento, na sua maioria, calculam uma medida de similaridade entre uma consulta q_i e um documento d_j, definindo uma ordem entre os documentos retornados em resposta a uma consulta. Mais especificamente, o cálculo da função de ranqueamento, normalmente, considera a importância ou peso de cada termo presente na consulta q_i em relação ao documento d_j. Ou seja, o cálculo da medida de similaridade para o documento d_j pode ser obtido a partir da soma dos pesos de cada termo presente na consulta para o documento d_j.

2.1.2.1 Modelo Booleano

No modelo booleano, a consulta é uma expressão booleana tradicional, que liga seus termos através de conectivos lógicos *AND*, *OR* E *NOT*. Neste modelo, um documento é considerado relevante ou não relevante a uma consulta, ou seja, não existe um resultado parcial e não há informação que permita a ordenação do resultado de uma consulta. Assim, esse modelo é geralmente utilizado para recuperação de dados do que para recuperação de informação [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013].

As principais vantagens deste modelo são o formalismo claro e a sua simplicidade, facilmente programável e exato. As desvantagens são: 1) os resultados podem possuir muitos documentos ou nulos, 2) os resultados não são ordenados. Apesar dessas desvantagens, o modelo booleano ainda é utilizando em sistemas comerciais [Baeza-Yates e Ribeiro-Neto 2013].

2.1.2.2 Modelo Vetorial

O modelo de recuperação de informação utilizado neste trabalho foi o modelo chamado de espaço vetorial e ele é o mais conhecido em recuperacão de informação

[Baeza-Yates e Ribeiro-Neto 2011]. A ideia central do modelo é representar algebricamente, como vetores em um espaco euclidiano n-dimensional, onde n é o número de termos distintos da coleção, o conjunto de termos de uma consulta q e dos documentos de uma coleção D. Um documento d_j é associado a um vetor $\vec{d_j}$ representado por $\vec{d_j} = (w_{1,j}, ..., w_{i,j}, ..., w_{t,j})$, onde $w_{i,j} \geq 0$ é o peso de um termo t_i em um documento d_j. Da mesma maneira, uma consulta q é associada a um vetor $\vec{q}$ representado por $\vec{q} = (w_{1,q}, ..., w_{i,q}, ..., w_{t,q})$, onde $w_{i,q} \geq 0$ é o peso de um termo t_i na consulta q.
Uma vez representados os documentos e a consulta em um espaco vetorial é possível calcular o grau de similaridade de um documento d_j em relação a uma consulta q como sendo a similaridade entre os vetores $\vec{d_j}$ e $\vec{q}$. Esta similaridade pode ser calculada, por exemplo, pelo cosseno do ângulo θ formado entre estes dois vetores. Esta fórmula é conhecida por medida do cosseno [Baeza-Yates e Ribeiro-Neto 2011] e está representada pela Equação 2.1.

$$sim(dj,q) = cos\theta = \frac{\vec{d_j} \times \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{t=} w_{t,j}^2 \times \sum_{t=1}^{t} w_{i,q}^2}} \tag{2.1}$$

onde $|\vec{d_j}|$ e $|\vec{q}|$ sao as normas dos vetores da consulta q e do documento d_j, sendo que o fator $|\vec{q}|$ não afeta o cálculo da similaridade, sendo geralmente simplificado, já que o valor de $|\vec{q}|$ é o mesmo para todos os documentos.
Existem diferentes formas para se calcular os pesos $w_{i,j}$ e $w_{i,q}$ [Salton e McGill 1986, Zobel e Moffat 1998], mas a estratégia mais utilizada é a conhecida por $tf \times idf$ [Baeza-Yates e Ribeiro-Neto 2011]. Onde tf (*term frequency*) é a frequência de um termo em um documento ou o número de vezes que um termo k_i ocorre em um documento d_j. Já o idf (*inverse document frequency*) é o inverso da frequência do documento ou o número de documentos nos quais um termo k_i é encontrado, considerando toda uma coleção de documentos. Dessa forma, uma maneira possível de se calcular os pesos $w_{i,j}$ e $w_{i,q}$ foi definida por Witten et al. [Witten et al. 1999], como mostra a Equação 2.2.

$$w_{i,j} = f(tf_{i,j}) \times idf + i = (1 + log tf_{i,j}) \times log(1 + \frac{N}{df_i}) \tag{2.2}$$

onde $tf_{i,j}$ é a frequência de um termo k_i em um documento d_j, N é o número de documentos da coleção e df_i é o número de documentos onde um termo k_i ocorre;

$$w_{i,q} = f(tf_{i,q}) \times idf + i = (1 + log tf_{i,q}) \times log(1 + \frac{N}{df_i}) \tag{2.3}$$

onde N é o número de documentos da coleção, df_i é o número de documentos onde um termo k_i ocorre e $tf_{i,q}$ é o número de ocorrências de um termo k_i em uma consulta q.

Existem muitas variações para $w_{i,j}$ e $w_{i,q}$ [Baeza-Yates e Ribeiro-Neto 2011]. Para $w_{i,j}$, por exemplo, a opçao é a definida na Equação 2.4.

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \tag{2.4}$$

onde $tf_{i,j}$ é a frequência de um termo k_i em um documento d_j, N é o número de documentos da coleção e df_i é o número de documentos onde um termo k_i ocorre.

2.1.2.3 Modelo Probabilístico

Este modelo foi introduzido por Robertson e Jones em 1976. A recuperacão deste modelo é vista como um problema de estimativa da probabilidade de que a representação de um documento corresponda ou satisfaça a representação de uma consulta. Deste modo, dado uma consulta de um determinado usuário, existe um conjunto de documentos que contém exatamente os documentos relevantes e nenhum documento irrelevante [Baeza-Yates e Ribeiro-Neto 2013].

Para determinar o conjunto de documentos relevantes, é necessário antes conhecer algumas características que definam este conjunto. Como essas características não são conhecidas no tempo de uma consulta, é necessário iniciar um processo específico para determiná-las [Baeza-Yates e Ribeiro-Neto 2013].

Seja uma consulta q e um documento d_j em uma coleção, o modelo probabilístico tenta estimar a probabilidade de um usuário considerar relevante o documento d_j. Este modelo assume que esta probabilidade de relevância depende apenas da representação da consulta e da representação do documento [Jones et al. 2000].

Esse modelo assume que há um subconjunto de documentos que o usuário prefere como resposta para a consulta q. Este conjunto de documentos ideais é representado por R e devem maximizar a probabilidade de relevância para o usuário. Documentos no conjunto R são rotulados como relevantes para a consulta q. Documentos que não estão neste conjunto são considerados não relevantes para q e são rotulados como $\overline{R}$, o complemento de R [Jones et al. 2000].

Os termos que ocorrem nos documentos em R podem ser utilizados para encontrar outros documentos relevantes. Este princípio é chamado de Princípio da Ordenação Probabilística e assume que a distribuição de termos na coleção seja capaz de informar a relevância provável de um documento para uma consulta [Baeza-Yates e Ribeiro-Neto 2013].

Neste modelo os termos são todos binários, por exemplo, $w_{i,j} \in 0, 1^2$ e $w_{i,q} \in 0, 1$. Uma consulta q é um subconjunto do índice de termos. Sejam R um conjunto de documentos relevantes, $\overline{R}$ o conjunto de documentos não relevantes, $P(R|\overrightarrow{d_j})$ a probabilidade do documento d_j ser relevante para a consulta q e $P(\overline{R}|\overrightarrow{d_j})$ a probabilidade do documento d_j não ser relevante para a consulta q. A similaridade

$sim(d_j, q)$ de um documento d_j em relacão a uma consulta q é definida na Equacão 2.5.

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\bar{R}|\vec{d_j})} \tag{2.5}$$

aplicando a regra de Bayes [Mitchell 1996], têm-se:

$$sim(d_j, q) = \frac{P(\vec{d_j}|R) \times P(R)}{P(\vec{d_j}|\bar{R}) \times P(\bar{R})} \tag{2.6}$$

onde $P(\vec{d_j}|R)$ é a probabilidade de selecionar aleatoriamente o documento d_j do conjunto R de documentos relevantes, $P(R)$ é a probabilidade de um documento selecionado aleatoriamente ser relevante, $P(\vec{d_j}|\bar{R})$ é a probabilidade de selecionar aleatoriamente o documento d_j do conjunto $\bar{R}$ de documentos não relevantes e $P(\bar{R})$ é a probabilidade de um documento selecionado aleatoriamente ser não relevante.

Sendo $P(R)$ e $P(\bar{R})$ os mesmos valores para todos os documentos da coleção, pode- se:

$$sim(d_j, q) \sim \frac{P(\vec{d_j}|R)}{P(\vec{d_j}|\bar{R})} \tag{2.7}$$

assumindo a independência entre os termos, tem-se:

$$sim(d_j, q) \sim \prod_{i=0}^{M} \frac{P(\vec{d_{j,i}}|R)}{P(\vec{d_{j,i}}|\bar{R})} \tag{2.8}$$

onde M é igual ao número de palavras distintas que ocorrem na coleção de documentos.

Uma vantagem deste modelo é o fato dos documentos serem ordenados de forma decrescente por suas probabilidades de serem relevantes. Este modelo é considerado, por alguns pesquisadores, melhor do que o modelo vetorial [Cooper 1994, Baeza-Yates e Ribeiro-Neto 2013].

As desvantagens deste modelo são:

- Separação aleatória da coleção de dois subconjuntos: documentos relevantes e documentos irrelevantes;
- Não faz uso da frequência dos termos no documento;
- Assume a independência entre termos.

Um exemplo de sucesso deste modelo é a função de ordenação BM25 [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013]. Esta função de ordenação faz uso da frequência dos termos no documento e também leva em consideracão o tamanho dos documentos, como pode ser visto na Equação 2.9.

$$sim(q,dj) = \sum_{t \in q} log(\frac{N}{n_i}) \times \frac{(k_1 + 1) \times f_{t,d}}{k((1 - b) + b \times (\frac{T_d}{T_{med}})) + f_{t,d}} \quad (2.9)$$

onde:

- $f_{t,d}$ é a frequência do termo t no documento d;
- T_d é o tamanho (quantidade de termos) do documento d;
- T_{med} é o tamanho médio dos documentos na coleção;
- k_1 é um parâmetro de valor positivo que calibra a escala do $d_{t,d}$. Se k_1 é igual a 0, então o retorno de $sim(q, d_j)$ é similar ao resultado do modelo booleano;
- b é um parâmetro ($0 \leq b \leq 1$) que determina a influência do tamanho dos documentos no cálculo da $sim(q, d_j)$. Quando $b = 0$ o valor de $sim(q, d_j)$ não é normalizado considerando T_d e T_{med}.

2.1.2.4 Resumo dos Modelos

Normalmente, o modelo booleano é considerado o modelo clássico mais falho. O problema principal em torno deste modelo é a incapacidade de reconhecer relevâncias parciais, documentos que satisfazem a necessidade do usuário de forma parcial. Essa característica acaba fazendo com que os sistemas que implementam este modelo tenham um baixo desempenho [Baeza-Yates e Ribeiro-Neto 2013].

Diversos trabalhos na literatura comparam os modelos vetorial e probabilístico [Salton e Buckley 1988, Croft e Lafferty 2003]. Alguns destes trabalhos indicam que o modelo probabilístico supera o modelo vetorial [Croft e Lafferty 2003] e outros sugerem que o modelo vetorial supera o modelo probabilístico em coleções de documentos genéricos [Salton e Buckley 1988]. O modelo vetorial é mais popular em implementações de Sistemas de Recuperação de Informacões [Manning et al. 2008].

Os três principais modelos de recuperação (Booleano, Vetorial e Probabilístico) estão no nível de palavras [Lugo e Alberto 2004]. Com isso, muitas pesquisas tentam responder à seguinte pergunta: a Recuperacão de Informação não melhoraria se fosse utilizadas técnicas mais sofisticadas para representar os documentos e as necessidades de informação do usuário? Muitos têm trilhado esta via, porém segundo [Russell e Norvig 2003] nenhuma das tentativas apresentou uma melhora significativa numa faixa ampla de tarefas relacionadas com a Recuperação de Informação.

2.1.3 Avaliação de Sistemas de Recuperação

As medidas mais utilizadas para avaliar o desempenho de um sistema computacional são tempo e espaço. Quanto menor o tempo de resposta de um sistema e quanto

menor o espaço em memória utilizado, melhor o sistema é considerado. Entretanto, para sistemas cujo objetivo é recuperar informações outras métricas devem ser utilizadas [Baeza-Yates e Ribeiro-Neto 2013].

Em uma consulta realizada por um usuário não existe uma resposta exata. Os documentos retornados por uma consulta são ordenados de acordo com a sua relevância em relação a consulta. As métricas utilizadas para avaliar um Sistema de Recuperação de Informação devem medir quão relevante é o conjunto de documentos, retornados pelo sistema, para o usuário [Baeza-Yates e Ribeiro-Neto 2013].

O surgimento de novas técnicas de Recuperação de Informação demandou como resultado a criação de novas técnicas de avaliação dos resultados. Para avaliar um sistema de Recuperação de Informação é fundamental estimar o quão bem o sistema atende às necessidades de informação do usuário. É difícil fazer esse cálculo, pois o mesmo conjunto resposta pode ter diversas interpretações por diferentes usuários. No entanto, pode ser feita definição de métricas aproximadas que, na média, podem ter uma relação entre as preferências de uma amostra e a populacão de usuários [Baeza-Yates e Ribeiro-Neto 2013]. Algumas destas métricas serão apresentadas na Seção 2.3.6.

Para saber como um sistema de recuperação de informacão está desempenhando é necessário uma avaliação adequada, que seja possível comparar os resultados recuperados com outros sistemas de recuperação de informação e responder questões que surgem na prática, tais como:

- É necessário fazer uma modificação na função de ranqueamento. Deve-se ir adiante e implementá-la?
- Uma nova função de ranqueamento foi implementada. Ela é superior a atual?

A avaliação da recuperação de informação é feita através de uma metodologia sistemática. Uma métrica quantitativa é associada aos resultados retornados pelo sistema de recuperação de informação em resposta a consultas feitas pelo usuário. Essa métrica está associada à relevância dos resultados para os usuários. É comum analisar a métrica pela comparação do resultado produzido pelo sistema de recuperação de informação com os resultados sugeridos por pessoas para o mesmo conjunto de consultas [Baeza-Yates e Ribeiro-Neto 2013].

2.1.3.1 Paradigma de *Cranfield*

A qualidade de um sistema de recuperação de informação é geralmente avaliada usando duas abordagens: pelo uso de gabaritos relacionados a consultas, ou usando observações de como um determinado usuário se comporta quando é apresentado um conjunto de resultados originados de uma consulta [Cleverdon 1991, Richardson et al. 2006, Zobel et al. 2011, Chapelle et al. 2012].

A primeira abordagem é baseada no chamado paradigma de *Cranfield* [Cleverdon 1991], que, apesar de antigo, é muito usado nos dias de hoje [Chapelle et al. 2012].

Em 1952, Cyril Cleverdon, um bibliotecário da escola de aeronáutica de Cranfield, Inglaterra, conheceu o sistema *Uniterm* proposto por Mortimer Taube, bibliotecário nos EUA. O sistema *Uniterm* possuía 40 mil títulos formados por 7 mil palavras distintas, e com o objetivo de organizar o seu ambiente de trabalho, Cleverdon indexou 200 documentos manualmente usando o sistema *Uniterm*. Em seguida solicitou que usuários fizessem diversas consultas. Cada consulta era baseada em um só documento e a busca era considerada bem-sucedida se aquele documento fosse localizado no catálogo.

A partir daí, Cleverdon propôs um estudo para comparar os diversos sistemas de indexação existentes. Ele obteve financiamento do *National Science Foundation* (NSF) para desenvolver o projeto que ficou conhecido como *Cranfield 1*, que envolveu a indexação manual de 18000 artigos sobre Engenharia Aeronáutica e avaliação de 1200 consultas.

Seis estudantes passaram três meses examinando cada documento em relação a consulta e decidindo quais documentos eram relevantes. O resultado foi uma coleção de teste composta por documentos, consultas e gabaritos para cada par consulta-documento. A coleção tornou-se conhecida como *Cranfiled 2*. Os experimentos Cranfield 2 estabeleceram a base para a experimentação moderna em Recuperação de Informação. O mesmo conjunto de documentos e consultas pode ser usado para avaliar sistemas de ranqueamento diferentes comparando-os com os gabaritos produzidos por especialistas humanos [Cleverdon 1991, Baeza-Yates e Ribeiro-Neto 2013].

Como vantagens da utilização deste paradigma, apresenta-se o custo relativamente baixo, a avaliação de um sistema de Recuperação de Informação pode ser feita rapidamente e seus resultados podem ser reproduzidos posteriormente para fins de verificação (repetibilidade). Além disso, pode ser aplicado focando tipos particulares de necessidades de informação (faces, plantas, imagens de satélite, imagens médicas e *Web*).

Os experimentos de Cleverdon culminaram em métricas modernas de avaliação da recuperação de informação, que serão abordadas na Seção 2.3.6. Para avaliar os resultados das consultas é importante saber a relevância dos objetos em relação à consulta. O conhecimento dos resultados dos gabaritos é fundamental, uma vez que para avaliação de um sistema de Recuperação de Informação, utilizando o paradigma de *Cranfield*, é necessário coleções de teste com 3 conjuntos fundamentais: um conjunto de documentos, um conjunto de consultas e um conjunto de gabaritos.

2.1.3.2 Métodos Alternativos ao Paradigma de *Cranfield*

Os métodos de avaliação da consulta baseada em usuários busca avaliar, de maneira adequada, a interface com o usuário e as interacões provocadas por estes. As preferências destes usuários podem ser afetadas facilmente por qualquer alteracão, seja nas características da interface ou pelo modo como o usuário vai expressar no sistema as suas preferências [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013]. Assim, alguns métodos são desenvolvidos especificamente para avaliar estes sistemas interativos, são eles: experimentação humana em laboratório, painéis lado a lado, teste A/B, crowdsoursing e avaliacão usando dados sobre cliques.

Experimentação humana em laboratório. Os participantes são selecionados criteriosamente e a avaliação é feita preferencialmente em ambiente fechado. São executadas várias sessões de testes, com diferenças mínimas na interface de teste, para que sejam avaliados os impactos destas alterações. As desvantagens deste método são alto custo e a limitação a um pequeno conjunto de necessidades de informação executados por um número limitado de usuários [Baeza-Yates e Ribeiro-Neto 2013].

Painéis lado a lado. Os testes são realizados por dois sistemas diferentes e o resultado dos testes são apresentados um ao lado do outro. O objetivo deste teste é permitir controlar as diferenças entre as opiniões dos participantes e as influências da opinião do usuário produzidas pelo ordenamento dos primeiros resultados. Uma das desvantagens deste método é que não é permitido avaliar se um sistema é melhor que o outro ou uma comparação direta entre vários sistemas, pois as condições de consulta em tempo real mudam com o tempo [Baeza-Yates e Ribeiro-Neto 2013].

Teste A/B. Este teste é realizado, por exemplo, com muitos usuários através de um chamamento por um site, onde a opinião dos mesmos é requisitada para avaliação da mudança no layout da página virtual. O objetivo é verificar como os usuários se comportam à mudança, sendo possível compreender se a modificação é positiva ou não. A quantidade de usuários que participam do teste constituem uma fração significativa dos usuários do site. Essa técnica é importante para sites que possuem muitos usuários, pois uma mudança considerada ruim pode causar causar problemas significativos entre os visitantes [Baeza-Yates e Ribeiro-Neto 2013].

Crowdsourcing. Termo usado para designar tarefa destinada a um grande grupo de pessoas, chamados de "trabalhadores", organizado em trabalhos colaborativos. Prevê que o usuário, a partir do aumento da interação com um grande número de usuários on-line, poderá trocar mensagens sugerindo questões que possam ser avaliadas quanto à relevância [Baeza-Yates e Ribeiro-Neto 2013].

Avaliação usando dados sobre cliques. Também chamado de *feedback* implí- cito, a avaliação com base na análise de dados sobre cliques tem sido utilizada e funciona por meio da observação da frequência com que os usuários clicam em um dado documento quando ele lhes é apresentado, com um baixo custo e sem trazer sobrecarga ao usuário. Este procedimento também requer preparação cuidadosa a fim de evitar falsos resultados [Baeza-Yates e Ribeiro-Neto 2013].

2.1.3.3 Coleções de Teste

Os experimentos de Cleverdon (1991) culminaram no paradigma de *Cranfield*, no qual são utilizadas coleções de teste. Essas coleções são frequentemente chamadas de carga de trabalho (*workload*) [Paramita et al. 2007], necessária para avaliar um sistema de recuperação de informação.

Para comparar a eficácia de um Sistema de Recuperação de Informação são usadas coleções de teste padronizadas. Uma coleção desse tipo consiste dos seguintes elementos [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013]: um conjunto de documentos; um conjunto de consultas, e; para cada consulta, um conjunto de julgamentos de relevância. Esse conjunto de julgamentos de relevância é identificado manualmente através de um processo que envolve um esforço humano significativo [Baeza-Yates e Ribeiro-Neto 2013]. Esse conjunto pode ser entendido através da função representada pela Equação 2.10.

$$rel(d,q) \in [0,1] \tag{2.10}$$

onde d é um documento do conjunto de documentos e q é uma consulta. De maneira geral, a Equação 2.10 é definida em um intervalo de 0 a 1. O valor 0 significa que o documento é completamente irrelevante para a consulta e 1 significa que o documento é completamente relevante para a mesma. Outros valores dentro desse intervalo representa o nível de relevância.

Conjunto de Documentos

Um conjunto de documentos pode conter alguns dados como título, autor e resumo ou então um texto completo. Informações adicionais podem ser utilizadas, tais como descritores designados por autor, um conjunto de termos usado como vocabulário de controle e informações sobre citacões.

Conjunto de Consultas

É fundamental obter consultas que reflitam necessidades reais e essa tarefa não é fácil, pois as necessidades de informação mudam em um curto período de tempo. Além disso, o tempo necessário para coletar consultas em relação ao conjunto de documentos pode ser longo. Surge então a pergunta: pode-se garantir que as consultas reais reflitam os documentos da coleção? Assim, existe a necessidade de criacão de consultas artificiais. No entanto, existe ainda um certo grau de dificuldade para retornar documentos específicos da coleção no momento da escolha das palavras-chave ou outros elementos formadores da consulta [Jones 1981].

A seleção de uma consulta é muito importante para uma coleção de teste, elas iniciam o processo avaliativo de um sistema de recuperação de informacão. Dessa forma, o conjunto de consultas deve explorar todos os aspectos do modelo de recuperaçao para permitir a generalização estatística dos resultados. Esse objetivo é

alcançado formando consultas combinadas com a maioria dos documentos da coleção [Jones 1981].

Conjunto de Julgamentos de Relevância ou Gabarito

Para cada consulta existente no conjunto de consultas são fornecidos documentos, pertencentes a coleção de documentos, considerados relevantes pelos usuários que submetem a consulta ou por especialistas do domínio. Para coleções grandes, geralmente são combinados os resultados de diferentes representações da consulta construída por diferentes usuários. A princípio, são preferidas as consultas e julgamentos obtidos diretamente dos usuários [Baeza-Yates e Ribeiro-Neto 2013].

A relevância está associada à necessidade de informação do usuário, ela não está associada a consulta. Um documento deve ser considerado relevante se e somente se suprir a necessidade de informação. Por exemplo, um documento não pode ser considerado relevante se todas as palavras que aparecem na consulta aparecem no documento também [Manning et al. 2008].

Geralmente, esse processo leva tempo e o custo é alto. Devido a isso é importante levar em consideração coleções de teste já existentes, que são usadas em pesquisas voltadas à área de recuperação de informacão e estão disponíveis online.

Coleções de Teste Existentes

Criar uma coleção de teste é um desafio e para que os experimentos alcancem níveis estatisticamente significativos, é importante que a coleção seja representativa. Neste sentido, é fundamental apresentar as coleções de teste existentes que são usadas em diferentes pesquisas na área de recuperação de informação e estão disponíveis [Sanderson 1994]. Essas coleções foram testadas e comparadas com outros sistemas de recuperação de informação, fornecendo resultados confiáveis.

Segundo Chapelle et al. (2012), os benefícios das coleções de teste existentes são:

- Coleções de teste podem ser usadas em pesquisas futuras. Isso significa que coleções de teste permitem que os experimentos sejam repetidos;
- A avaliação de um sistema de recuperação de informacão é feita mais rapidamente, por exemplo, para comparação de sistemas com variações de uma determinada função de ranqueamento;
- Diferentes coleções de teste podem ser desenvolvidas com o foco de tipos particulares de necessidade de informação, permitindo, assim, um melhor entendimento e validação do comportamento da função de ranqueamento.

Atualmente existem inúmeras iniciativas de apoio e fomento à pesquisa, campanhas de avaliação de sistemas, desafios de pesquisa, etc. relacionadas à área de RI, tais como: CLEF, NTCIR, OHSUMED, Reuters, INEX, entre outras.

1. CLEF [1]. É uma conferência que ocorre anualmente que tem por objetivo tratar de assuntos relacionados a Recuperação de Informação Multilíngue e Multimodali-

[1]CLEF - *Workshop on Cros-language IR and Evaluation*

dade. As coleções disponibilizadas por essa iniciativa dão suporte a experimentos relacionados a filtragem de informacões multilíngues, propriedade intelectual, recuperaçao de vídeo *cross-language*, dentre outros [Müller et al. 2010]. Em 2005, a CLEF contou com uma campanha de *benchmark*, chamada de GeoCLEF[2], considerada a primeira conferência de avaliacão de um sistema de recuperação geográfica, na qual forneceu um *framework* para avaliar sistemas de recuperação geográfica analisando aspectos espaciais e multil´ıngues [Gey et al. 2005]. O desafio da GeoCLEF era encontrar documentos relevantes de coleções textuais. Exemplo de uma necessidade de informação geográfica da GeoCLEF: encontrar histórias sobre desastres em Genebra. Genebra é uma cidade da Suíça e, portanto, é considerada uma informação geográfica. No entanto, sem a informação de latitude e longitude necessária para este trabalho.

2. GikiP[3]. Uma tarefa chamada *tarefa Wikipedia* que se concentra na informacão geográfica na Wikipédia em um ambiente multilíngue [Cardoso 2007].

3. NTCIR (*NII Test Collection for IR Systems*). As coleções de teste geradas nos Workshops dessa organização são compostas por patentes em diversos idiomas. Essas coleções permitem realizar experimentos na área de tradução de patentes, recuperação de dados em patentes e recuperação multilíngue. Os workshops são promovidos anualmente para fomentar a pesquisa na área de Recuperação de Informação, sumarização de texto, perguntas e respostas, entre outras. [Baeza-Yates e Ribeiro-Neto 2013].

4. INEX [4]. Estas coleções de teste são para experimentos relacionados a recuperação de informação sobre dados do tipo XML. Portanto, é de suma importância para a área de recuperação de informação e também para a comunidade de XML.

5. *Reuters-21578* . A *Reuters-21578* é uma das coleções mais utilizadas para realização de experimentos em recuperacão de informações. Os artigos dessa coleção são da agência de notícias *Reuters* [Lewis 2004]. Possui 21578 documentos e é o sucessor da coleçao *Reuters-22173* [Sanderson 1994]. Essa coleção possui aproximadamente 20Mb de tamanho e não fornece consultas para avaliacão. Uma abordagem de criar tais consultas *Reuters-21578* foi proposta por [Sanderson 1994] e está dispon´ıvel para *download* no endereço https://goo.gl/NrOfu.

6. NewsGroups, OHSUMED. A coleção NewsGroups (20-NewsGroups) possui um conjunto de aproximadamente 20.000 mensagens que foram postadas nos grupos de notícias da *Usenet*. Por sua vez, a OHSUMED é um banco de dados composto pela literatura médica mantida pela NLM [5], composta por 348.566 referências mé- dicas selecionadas entre os anos de 1987-1991 [Hersh et al. 1994].

[2]GeoCLEF - http://www.clef-campaign.org/

[3]GikiP - http://www.linguateca.pt/GikiP/

[4]INEX - *INitiative for the Evaluation of XML Retrieval.*

[5]NLM - *Nacional Library Medicine*

2.1.3.4 Métricas de Avaliação

As métricas de avaliação quantificam a similaridade entre o conjunto de documentos recuperados e o conjunto de documentos considerados relevantes pelos especialistas. Isto fornece uma estimativa da qualidade do Sistema de Recuperação de Informação avaliado. As métricas utilizadas para a avaliação dos Sistemas de Recuperação de Informação são uma maneira de quantificar algo inerentemente subjetivo [Manning et al. 2008, Baeza-Yates e Ribeiro-Neto 2013].

Existem inúmeras métricas que podem ser utilizadas para avaliar o desempenho de um Sistema de Recuperação de Informação. As principais são Precisão e Revocação [Baeza-Yates e Ribeiro-Neto 2013].

Levando-se em consideração um conjunto de necessidades de informação, seja R o conjunto de documentos relevantes e A o conjunto-resposta gerado pela execução de uma consulta que processa a necessidade de informação [Baeza-Yates e Ribeiro-Neto 2013].

Revocação. Revocação ou também chamado de Sensibilidade (*Sensitivity*) é a fração do número de documentos relevantes recuperados pelo número total de documentos relevantes que existem na coleção [Baeza-Yates e Ribeiro-Neto 2013].

$$Revocação = \frac{|R \cap A|}{|R|} \tag{2.11}$$

Precisão. Precisão é a fração de documentos relevantes recuperados em relação ao número de documentos recuperados [Baeza-Yates e Ribeiro-Neto 2013].

$$Precisão = \frac{|R \cap A|}{|A|} \tag{2.12}$$

Medidas-F e F_1. Um Sistema de Recuperação de Informacão pode equilibrar Precisão e Revocação. No caso extremo, um sistema que retorna todos os documentos de uma determinada coleção de documentos como seu conjunto resposta, garantindo uma cobertura igual a 100%, no entanto, terá uma baixa Precisão. Em contrapartida, um sistema pode retornar um único documento e ter um baixo índice de cobertura, mas teria uma chance razoável de 100% de Precisão [Russell e Norvig 2003]. Uma maneira de definir uma média harmônica entre Precisão e Revocação é através da medida F [Manning et al. 2008].

As métricas Medida-F e F_1 combinam os valores de precisão e revocação em um único valor, permitindo atribuir diferentes pesos a elas [Baeza-Yates e Ribeiro-Neto 2013].

A definição formal pode ser vista na Equacão 2.13.

$$MedidaF_\alpha = \frac{(\alpha^2 + 1) \times Precisão \times Revocação}{\alpha^2 \times Precisão + Revocação} \tag{2.13}$$

onde o valor de α define a importância relativa de precisão e revocação. Se o $\alpha = 0$, somente a precisão é considerada. Quando o $\alpha = \infty$, somente a revocação é considerada.

Para a medida-F_1, considera-se o $\alpha = 1$. Com esse valor, precisão e revocação têm pesos iguais.

Precisão em k **(**$P@k$**).** A $P@k$ mede a relevância dos k primeiros documentos de uma lista ordenada. A Equação 2.14 ilustra como é feito o cálculo da $P@k$.

$$P@k = \frac{r}{n} \tag{2.14}$$

onde n é o número de documentos retornados e r é o número de documentos considerados relevantes e retornados até a posição n da lista ordenada. Por exemplo, se os 10 primeiros documentos retornados por uma consulta são *relevante, relevante, irrelevante, relevante, relevante, relevante, irrelevante, irrelevante, relevante, rele- vante* então os valores de $P@1$ até $P@10$ são 1, 1, 2/3, 3/4, 4/5, 5/6, 5/7, 5/8, 6/9, 7/10, respectivamente. Para um conjunto de consultas, deve-se calcular a média de $P@k$.

***Mean Average Precision* (MAP).** MAP é uma métrica que agrupa todos os valores de $P@k$. *Average Precision* (AP) é definido como uma média para todos os valores de $P@k$, como pode ser visto pela Equacão 2.15.

$$AP = \frac{\sum_{n=1}^{N} P@k \times rel(n)}{r_q} \tag{2.15}$$

onde r_q é o número total de documentos considerados relevantes para a consulta; N é o número de documentos recuperados, e; rel(n) é uma função binária sobre a relevância do n^{th} documento:

$$rel(n) = \begin{cases} 1 & \text{se o } n^{th} \text{ documento é relevante} \\ 0 & \text{caso contrário} \end{cases} \tag{2.16}$$

Assim sendo, o valor de MAP é a média dos valores de AP para todas as consultas.

2.2 Consulta Espacial Por Palavra-Chave

Com o aumento de objetos *online* com informação textual e espacial (latitude e longitude), a internet torna-se um ambiente dimensionalmente espacial [Chen et al. 2013]. Os usuários com *smartphones*, *tablets* e outros aparelhos com GPS (*Global Positioning System*) e os conteúdos da *Web* estão cada vez mais geoposicionados e geo codificados. Além disso, pontos de interesse estão cada vez mais dispon´ıveis na *Web*.

Esse desenvolvimento necessita de técnicas que permitam a indexação de dados que contenham descrições textuais e espaciais. Uma dessas técnicas é a consulta espacial por palavra-chave que tem como argumentos a localização espacial e um conjunto de palavras-chave, retornando conteúdo levando em consideração a informação textual e a localização espacial [Cao et al. 2012].

São três os tipos de consulta espacial por palavra-chave mais importantes: A consulta k NN (k-*Nearest-Neighbor*) booleana, a consulta *range* booleana e a consulta k NN top-k.

- **Consulta kNN Booleana**. Retorna os k objetos mais próximos à localização da consulta (representada por um ponto), tal que cada descrição textual contém todas as palavras-chave da consulta (AND). A Figura 2.2 representa um exemplo da consulta, que possui três argumentos: a localização da consulta q, o conjunto de palavras chave ("***bicicleta***", "***esporte***") e o número de resultados esperados ($k = 2$). Os dois objetos mais próximos e que obedecem os critérios da consulta kNN booleana são o objeto p_4 e o objeto p_5 [Chen et al. 2013]. Observa-se que o objeto p_3 não é relevante por não possuir todas as palavras-chave da consulta;

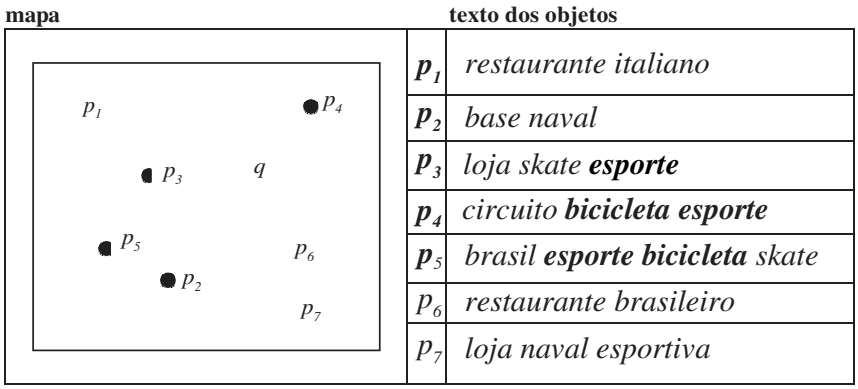


Figura 2.2: Consulta k NN booleana.

- **Consulta k NN Top-k** . Retorna os k objetos ordenados por um *ranking* levando em consideração a distância dos objetos em relação à localização da consulta e a relevância textual dos objetos em relação às palavras-chave da consulta. Essa consulta possui o mesmo número de argumentos que a consulta k NN booleana, no entanto, os critérios de ranqueamento são definidos levando em consideração a proximidade espacial e a relevância textual. A Figura 2.2 representa um exemplo da consulta kNN top-k onde são retornados os objetos p_3 e p_4 [Chen et al. 2013];
- **Consulta *Range* Booleana**. Retorna todos os objetos cuja descrição textual contém as palavras-chave da consulta e cuja localização fica a menos de uma determinanda distância (especificada pelo usuário - *range*) do local da consulta.

A Figura 2.3 representa um exemplo onde a consulta q possui como argumentos: *range* de 2 quilômetros da consulta q e as palavras-chave "bicicleta" e "esporte". O resultado desta consulta são os objetos p_3 e p_5 [Chen et al. 2013].

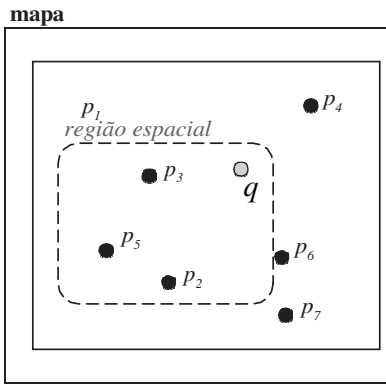


Figura 2.3: Consulta *Range* Booleana.

2.2.1 Consulta Espaço-Textual Top-*k*

A consulta espaço-textual top-k é uma consulta espacial por palavra-chave do tipo kNN Top-k [Chen et al. 2013], onde utiliza-se a localização do usuário, um conjunto de palavras-chave, fornecidas por ele, e o número de resultados como parâmetros. A consulta identifica objetos que são espacialmente próximos à localização do usuário, e textualmente relevantes às palavras-chave, retornando os k objetos que melhor atendem a estas duas características (proximidade espacial e relevância textual). Uma função de ranqueamento avalia a proximidade espacial entre um objeto e o usuário, além da relevância textual da descrição do objeto considerando o conjunto de palavras-chave. A resposta desta consulta é ordenada a partir dos valores de escore gerados para cada objeto pela função de ranqueamento [Rocha-Junior et al. 2011, Cao et al. 2012, Chen et al. 2013]. Uma das funções de ranqueamento utilizada na execução de uma consulta espaço-textual top-k é a função representada pela Equação 2.17 (EQA) [Cong et al. 2009].

$$rank(p,q) = \alpha \times \delta(p.l, q.l) + (1-\alpha) \times \theta(p.d, q.d) \tag{2.17}$$

onde p é o objeto, q é a consulta, $p.l$ é a localização do objeto, $q.l$ é a localização da consulta, $p.d$ é o texto vinculado ao objeto e $q.d$ representa as palavras-chave vinculadas à consulta.

Além da função de ranqueamento representada pela Equação 2.17, uma outra função de ranqueamento pode ser utilizada para uma Consulta Espaço Textual Top-k, representada pela Equação 2.18 (EQB) [Rocha-Junior e Nørvåg 2012].

$$rank(p,q) = \frac{\theta(p.d,\ q.d)}{1 + \alpha \times \delta(p.l, q.l)} \tag{2.18}$$

A função interna $\delta(p.l, q.l)$ calcula a proximidade espacial entre a localização da consulta e a localização do objeto. A função interna $\theta(p.d, q.d)$ calcula a relevância textual entre o texto vinculado ao objeto e as palavras-chave vinculadas à consulta. Tanto a função de proximidade espacial quanto a função de relevância textual retornam valores entre 0 e 1. O parâmetro de balanceamento $\alpha \in [0, 1]$ é utilizado para definir a importância de um critério (proximidade espacial ou relevância textual) sobre a outra [Rocha-Junior 2012].

Proximidade Espacial (δ). O cálculo da proximidade espacial é calculado através da Equação 2.19. A Equação 2.17 normaliza o valor da distância entre a localização do objeto e a localização da consulta. O d_{max} é a maior distância que dois pontos podem possuir no espaço.

$$\delta(p.l,\ q.l) = 1 - \frac{d(p.l,\ q.l)}{d_{max}} \tag{2.19}$$

Relevância Textual (θ). A relevância textual das duas equações pode ser computada utilizando a Equação 2.20 [Cong et al. 2009]. A função adota uma conhecida maneira de calcular a relevância textual, chamada de cosseno [Baeza-Yates e Ribeiro-Neto 2011].

$$\theta(p.d,q.d) = \frac{\sum_{t \in q.d} w_{t,p.d} \times w_{t,q.d}}{\sqrt{\sum_{t \in p.d} (w_{t,p.d})^2 \times \sum_{t \in q.d} (w_{t,q.d})^2}} \tag{2.20}$$

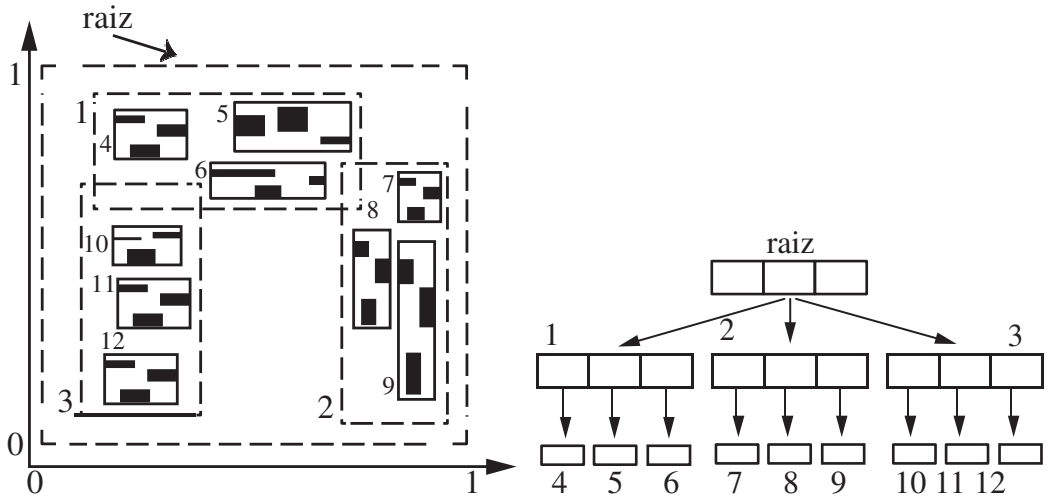
onde $w_{t,p.d}$ representa o peso do termo t no objeto $p.d$. A Equação 2.20 calcula o cosseno entre os dois vetores ($w_{t,p.d}$ e $w_{t,q.d}$), de forma que quanto mais próximo de 1 seja o cosseno, mais similares são os objetos.

2.2.2 Índices Espaço-Textuais

Os sistemas de processamento de consultas espaciais por palavra-chave atingem melhor eficiência quando são utilizados índices de indexação que combinam índices textuais e espaciais, chamados assim de índices híbridos. Esses índices podem ser categorizados em: índices baseados em R-*tree* [Zhou et al. 2005] [Hariharan et al. 2007] [Cary et al. 2010] [Rocha-Junior et al. 2011], índices baseados em *Grid* [Vaid et al. 2005] [Khodaei et al. 2010] e índices baseados na curva de preenchimento espacial [Chen et al. 2006] [Yan et al. 2009] [Christoforaki et al. 2011].

2.2.2.1 Índices Espaciais Baseados em R-*tree*

O R-*tree* é representado espacialmente pelo seu Retângulo Envolvente Mínimo (*Minimum Bounding Rectangle*- MBR). Os nós pais são formados por retângulos de dados, que são agrupados recursivamente para formar os nós avós, produzindo uma estrutura de árvore. O MBR do pai contém os MBRs de seus filhos e os MBRs podem se sobrepor. Cada nó da árvore formado representa uma página de disco, que são posições sucessivas de bytes na superfície do disco que são recuperados com um acesso a disco. A Figura 2.4(a) ilustra os retângulos de dados (em preto), organizados em uma R-*tree*. A Figura 2.4 (b) ilustra a estrutura do arquivo para a mesma R-*tree*, no qual os nós correspondem a páginas de disco.



(a) Representação espacial dos nós de uma R-*tree*.

(b) A estrutura de arquivo da R-tree.

Figura 2.4: Estrutura de uma R-*tree*. (adaptado de [Baeza-Yates e Ribeiro-Neto 2013]).

Diversos são os índices desta categoria: os índices "Arquivo Invertido R*-*tree*" (IF-R*) e "R*-*tree* Arquivo Invertido" (R*-IF) são exemplos de índices espaco-textuais que combinam os índices R-*tree* e Arquivo Invertido. Ambos são utilizados nas consultas *range* booleana, no entanto, o IF-R* possui um desempenho melhor [Zhou et al. 2005]; o KR*-*tree* (*Keyword* R*-*tree*) é outro índice espaço-textual pro- posto para utilização em consultas *range* booleana. Esse índice captura a distribuição de palavras-chave no espaço, melhorando o desempenho [Hariharan et al. 2007]; o IR^2-*tree* é um índice espaço-textual que pode ser usado para as consultas *range* booleana e as consultas *k* NN booleana, este ´ındice integra arquivo assinado em cada nó da R-*tree* [Felipe et al. 1984]; por fim, o índice *Spatial Inverted Index* (S2I) [Rocha-Junior et al. 2011], que também é baseado no R-*tree*, foi proposto inicialmente para consultas *k* NN top-*k*, no entanto, pode ser usado para as consultas *range* booleana e para consultas *k* NN booleana.

O *Spatial Inverted Index* (S2I) é um índice espaco-textual projetado para consultas espaço-textuais top-k [Rocha-Junior 2012], mas pode ser empregado em consultas kNN booleanas e consultas range booleanas [Chen et al. 2013]. Esse índice organiza a frequência dos termos da seguinte maneira: os termos com maior frequência são armezados em aR-*trees* [6], uma árvore por termo; os termos com menor frequência são armazenados em blocos em um arquivo invertido.

Os três componentes do S2I são:

- **Vocabulário**. Esse componente armazena, para cada termo diferente, o número de objetos que esse termo aparece e um *flag* indica que tipo de estrutura é usada por cada termo (bloco ou *tree*).
- **Bloco**. Cada bloco armazena um conjunto de objetos. Cada objeto possui um *id* de identificação, a sua localização e o impacto do termo na coleção. Os objetos armazenados no bloco não são ordenados.
- **Árvores**. Uma aR-*tree* segue a mesma estrutura de uma R-*tree* tradicional. Um nó folha armazena a informação dos objetos (identificação, localização e peso). Todos os objetos de uma sub-árvore são armazenados em um retângulo envolvente mínimo (*MBR*. Diferentemente de uma R-*tree* tradicional, as entradas de uma aR-*tree* também armazenam valores não espaciais entre os objetos e suas sub-árvores.

2.2.2.2 Índices Baseados em *Grid*

Nesta categoria, combina-se um índice *grid* com um índice textual (arquivo Invertido por exemplo). Os índices de *grid* dividem o espaço em células de igual tamanho e as partes textuais ou em grid podem ser organizadas juntas ou separadas [Vaid et al. 2005] [Khodaei et al. 2010].

Em 2005, Vaid et al. realizaram um estudo sobre indexação espaco-textual em um engenho de busca geográfico na *Web* e propuseram dois índices espaço-textuais:

- **Índice de Prioridade Espacial**. Este índice prioriza a dimensão espacial na estrutura de indexação e acesso ao índice. A cobertura imposta, pelos locais de interesse encontrados nos documentos, corresponde ao espaço geográfico que é dividido em células de igual tamanho, formando um *grid*. Um Arquivo Invertido é construído para cada célula funcionando como um índice "Puramente Textual", contendo documentos cujo conjunto de locais de interesse faz uma interseção com a célula espacial;
- **Índice Espaço-Textual de Prioridade Textual**: o funcionamento deste índice é inverso ao anterior, priorizando a dimensão textual. O modelo

[6] *aggregate* R-*tree* - proposto por [Papadias et al. 2001] para operações *OLAP* eficientes, melho- rando o processamento da consulta espaço-textual por palavra-chave.

> "Puramente Textual" é alterado de maneira que, para cada conjunto de documentos vinculados a um termo do ´ındice, este seja separado em grupos organizados espacialmente. Isso significa que as referências aos documentos serão organizadas da seguinte maneira: $Celula_1(ConjuntodeDocumentos_1)$; $Celula_2(ConjuntodeDocumentos_2)$; $Celula_3(ConjuntodeDocumentos_3)$. Cada um destes conjuntos contém os documentos cujo conjunto de locais de interesse faz a interseção à célula associada.

Em 2010, Ali Khodaei et al. propuseram uma estrutura de arquivo invertido por palavra-chave espacial (*Spatial-Keyword Inverted File*) com o objetivo de armaze- nar informação espacial e textual para que objetos espaciais e textuais possam ser utilizados simultaneamente. Para a informação espacial este índice assume que cada objeto tem uma região contendo sua localização e para informação textual é utilizado o ´ındice Arquivo Invertido.

2.2.2.3 Índices Baseados na Curva de Preenchimento Espacial

Uma curva de prenchimento espacial é um objeto geométrico fractal[7] [Sagan 1994]. Este objeto preenche todo o espaço no intuito de mapear pontos de um determinado espaço multidimensional em uma só dimensão. A Figura 2.5 apresenta três curvas bidimensionais de preenchimento espacial: *Hilbert* [Hilbert 1981], *Lebesgue* ou "Z-*order* " e a de *Sierpinski*.

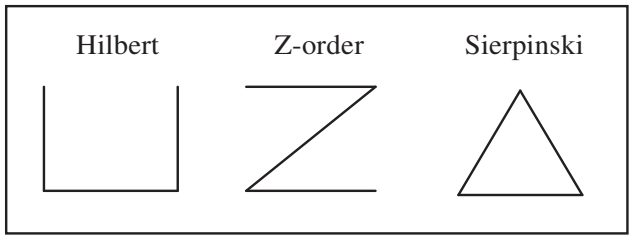


Figura 2.5: Curvas de *Hilbert*, de "Z-*order* " e de *Sierpinski* (adaptado de [Hilbert 1981]).

As curvas preservam as relações de ordem entre a proximidade dos pontos no espaço. Os pontos que são mapeados próximos à curva tendem a corresponder a pontos próximos no espaço.

Em 2006, Yen-Yu Chen [Chen et al. 2006] combinou "Arquivo Invertido" com a curva de *Hilbert* no intuito de otimizar o processo de indexação, neste trabalho o objeto espaço-textual tem uma região que representa um mapa de estradas ou um mapa com múltiplas localizações. Em 2011, Christoforaki et al. propuseram índices que combinam a curva de preenchimento espacial e Arquivo Invertido. Dentre os

[7]Figuras da geometria não-Euclidiana

índices propostos, o SFC-QUAD se mostrou mais eficiente que os demais. Este índice foi inicialmente projetado para consultas *range* booleanas, no entanto, podem ser usadas para as consultas k NN booleanas e para as consultas k NN top-k.

2.3 Considerações Finais

Foi apresentado neste capítulo a fundamentacão teórica relacionada a esta pesquisa. Dessa forma, foi apresentado conceitos da área de Recuperação de Informação: arquivo invertido, modelos de recuperação de informação (Booleano, Vetorial e Probabilístico), avaliação de sistemas de recuperação de informação (Paradigma de Cranfield, Métodos alternativos ao paradigma de Cranfield, Coleções de Teste e Métricas de avaliação), Consulta espacial por palavra-chave (Consulta Espaço-Textual Top-k, Índices espaço-textuais (Baseados em R-*tree*, *Grid* e Curva de Preenchimento Espacial).

Nesta pesquisa, as consultas utilizadas nos experimentos são do tipo vetorial. Além disso, o a avaliação dos resultados é feita utilizando o Paradigma de Cranfield, que estabelece principalmente que um sistema deve ser avaliado utilizando coleções de teste. Especificamente neste mestrado, as coleções de teste são formadas por um conjunto de documentos espaço-textuais, uma consulta e um conjunto de julgamentos de relevância. Para quantificar a qualidade, foram aplicadas as métricas Precisão (para avaliar qualitativamente a consulta quanto à relevância), a *Average Spatial Similarity* (que será abordada no próximo Capítulo, utilizada para avaliar a qualidade do resultado quanto à distância) e a Média Harmônica F_1.

No Capítulo a seguir é apresentada uma proposta de metodologia de avaliação da Consulta Espaço-Textual Top-k.

Capítulo 3

Proposta de Metodologia de Avaliação

"Os que desprezam os pequenos acontecimentos nunca farão grandes descobertas. Pequenos momentos mudam grandes rotas"

– Augusto Cury

Este capítulo apresenta a proposta utilizada nesta pesquisa para avaliar qualitativamente a Consulta Espaço-Textual Top-k, com as etapas necessárias e em que ordem essas etapas devem ser seguidas até a realização de testes e a aplicação das métricas para avaliar os resultados obtidos em seu contexto textual e espacial.

O "quadro geral" de avaliação é apresentado através de um diagrama de atividades, como mostra a Figura 3.1. As próximas seções detalham cada etapa apresentada na figura. A Seção 3.1 apresenta a fase de preparação, que consiste em preparar os dados para a realização dos experimentos. A Seção 3.2 apresenta o procedimento necessário para realização dos experimentos. Finalmente, a fase de análise é apresentada na Seção 3.3.

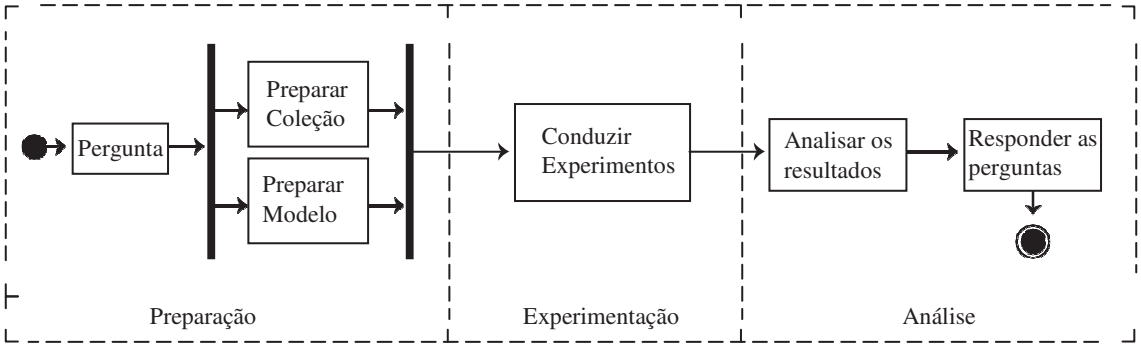


Figura 3.1: Etapas de Avaliação.

3.1 Fase de Preparação

A fase de preparação envolve os componentes necessários para a realização de uma avaliação adequada. Isso inclui, preparar uma coleção de teste para que o sistema seja testado, bem como escolher um modelo de recuperação de informacão que execute as consultas nos experimentos. Portanto, esta seção apresenta a maneira utilizada neste trabalho para preparar a coleção de referência espaço-textual a partir de uma coleção de teste tradicional.

Segundo Baeza Yates (2013) [Baeza-Yates e Ribeiro-Neto 2013], uma coleção de teste deve possuir um conjunto de documentos, um conjunto de consultas e um conjunto de julgamentos de relevância para cada par consulta-documento. No que diz respeito à avaliação da Consulta Espaco-Textual Top-k, cada documento do conjunto de documentos, de uma coleção de teste, deve possuir um texto e a sua localização (latitude e longitude). Esta seção apresenta o procedimento utilizado neste trabalho para preparar uma coleção de teste espaço-textual a partir de uma coleção tradicional.

A coleção escolhida foi a Reuters-21578, que é considerada uma das principais referências para avaliação de sistemas de recuperação de informação textual. É uma coleção de 21578 documentos, originalmente coletados e rotulados pela agência de notícias *Reuters*[1] no ano de 1987. Os documentos são agrupados em cinco categorias principais (*Exchanges*, *People*, *Topics*, *Organizations* e *Places*), que são divididas em subcategorias [Manning et al. 2008].

Os documentos desta coleção estão no formato XML (eXtensible Markup Language), com as seguintes *tags*: *date*, *topics*, *places*, *people*, *orgs*, *exchanges*, *companies*, *unk- nown* e *text*. As consultas são extraídas das *tags topics*, *people*, *orgs*, *places* e *ex- changes*, que representam as categorias principais da coleção Reuters-21578.

Um documento típico desta coleção pode ser visto na Figura 3.2. Esse documento pertence às categorias *Topics* e *Places*. Assim como este, todos os outros documentos

[1] www.reuters.com

da coleçao não possuem uma localização espacial (latitude e longitude). Devido a isto, essa coleção não é propícia para avaliação de sistemas de recuperação espaço-textuais, que necessitam de um conjunto de documentos com informacões textuais e espaciais. Assim, para que esta base possa ser utilizada na avaliação qualitativa da Consulta Espaço-Textual Top-k, é necessário que seus documentos possuam uma localização espacial. A forma como a localização espacial é incluída é importante e é discutida em detalhes na Seção 3.1.3.

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="19428" NEWID="3010">
<DATE> 9-MAR-1987 08:13:36.29</DATE>
<**TOPICS**><D>**strategic-metal**</D></**TOPICS**>
<**PLACES**><D>**usa**</D><D>**south-africa**</D></**PLACES**>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN> &#5;&#5;&#5;A &#22;&#22;&#1;f0808&#31;reute
d f AM-URANIUM 03-09 0135
</UNKNOWN>
<TEXT>&#2;
<TITLE>U.S. TO ALLOW TEMPORARY IMPORTS OF S.A. URANIUM</TITLE>
<DATELINE> WASHINGTON, March 9 - </DATELINE><BODY>The Treasury
Department said it would temporarily permit imports of South African
uranium ore and oxide pending clarification of anti-apartheid
sanctions laws passed by Congress last fall. The decision was announced
late Friday. It applies, until July 1, to uranium ore and oxide imported into
the U.S. for processing and re-export to third countries. The Treasury
said it took the action because it felt that when Congress passed the
comprehensive South African sanctions bill last fall over President
Reagan's veto it had not intended to hurt U.S. industry. In addition, the
Treasury said it would permit U.S.-made goods to be imported
temporarily from South African state-controlled organizations for repair
or servicing. Reuter &#3;</BODY></TEXT>
</REUTERS>

Figura 3.2: Exemplo de um documento da coleção Reuters-21578.

3.1.1 Seleção das Consultas

A coleção Reuters-21578 não possui um documento descrevendo um conjunto de consultas e os documentos que são relevantes para esta consulta. O mapeamento entre consulta e documentos relevantes para a consulta é importante para a avaliação qualitativa da Consulta Espaço-Textual Top-k. Para resolver estes problemas, optou-se por extrair consultas a partir das subcategorias das categorias principais. Essas subcategorias são consideradas palavras-chave da consulta, pois se um documento

da coleção está em uma determinada subcategoria, presume-se que este documento é relevante para uma consulta com esta palavra-chave.

Para exemplificar como as consultas são selecionadas, a Figura 3.2 apresenta um documento típico da coleção Reuters-21578, que pertence às subcategorias ***strategic-metal*** , ***usa*** e ***south-africa*** . Isto posto, se um usuário fizer uma consulta com as palavras-chave "strategic-metal", "usa" e "south-africa", o documento da Figura 3.2 é considerado relevante para esta consulta.

Nesta pesquisa as consultas foram agrupadas por número de subcategorias que um documento pertence. Desse modo, foram criados grupos de consultas com uma, duas, três e quatro palavras-chave. Outrossim, um documento só será relevante para uma determinada consulta se pertencer às subcategorias relacionadas às suas palavras-chave.

Por exemplo, a Figura 3.2 apresenta um documento pertecente às subcategorias ***strategic-metal*** , ***usa*** e ***south-africa*** , portanto, esse documento é relevante para qualquer consulta com a combinação dessas palavras-chave. Ao passo que, caso a consulta possua as palavras-chave "strategic-metal", "usa", "south-africa" e "poehl" o documento da Figura 3.2 não é relevante para essa consulta, no entanto, é relevante para o documento da Figura 3.3.

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="19429" NEWID="3011">
<DATE> 9-MAR-1987 08:13:36.29</DATE>
<**TOPICS**><D>**strategic-metal**</D></**TOPICS**>
<**PLACES**><D>**usa**</D><D>**south-africa**</D></**PLACES**>
<**PEOPLE**><D>**poehl**</D></**PEOPLE**>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN> &#5;&#5;&#5;A &#22;&#22;&#1;f0808&#31;reute
d f AM-URANIUM 03-09 0135
</UNKNOWN>
<TEXT>&#2;
<TITLE>U.S. TO ALLOW TEMPORARY IMPORTS OF S.A. URANIUM</TITLE>
<DATELINE> WASHINGTON, March 9 - </DATELINE><BODY>The Treasury
Department said it would temporarily permit imports of South African
uranium ore and oxide pending clarification of anti-apartheid
sanctions laws passed by Congress last fall. The decision was announced
late Friday. It applies, until July 1, to uranium ore and oxide imported into
the U.S. for processing and re-export to third countries. The Treasury
said it took the action because it felt that when Congress passed the
comprehensive South African sanctions bill last fall over President
Reagan's veto it had not intended to hurt U.S. industry. In addition, the
Treasury said it would permit U.S.-made goods to be imported
temporarily from South African state-controlled organizations for repair
or servicing. Reuter &#3;</BODY></TEXT>
</REUTERS>

Figura 3.3: Documento pertencente às subcategorias **strategic-metal**, **usa**, **south-africa** e **poehl**.

3.1.2 Seleção dos Gabaritos

Em uma coleção de teste, é importante que exista um conjunto de consultas e, para cada consulta, um conjunto de julgamentos de relevância (Gabaritos). A Coleção Reuters-21578 não possui esse conjunto e esta seção apresenta os critérios utilizados nesta pesquisa para criar gabaritos para cada consulta selecionada.

Um gabarito representa os documentos relevantes de uma determinada consulta ou o conjunto de julgamentos de relevância. Neste trabalho, após a seleção das consultas é feito um gabarito para cada uma. Primeiro foram estabelecidos os critérios de relevância. Assim, um documento é considerado relevante se ao menos uma das palavras-chave da consulta estiver presente no texto das categorias. Por exemplo, na consulta cuja palavra-chave é "livestock", todos os documentos que tiver a palavra "livestock" no texto das categorias é considerado relevante.

Para o grupo de consultas pertencentes a duas palavras-chave, os documentos que possuem estas palavras-chave são relevantes para essas consultas. Por exemplo, na consulta com as palavras-chave "strategic-metal" e "usa", todos os documentos que possuem estes termos no texto das categorias são considerados relevantes.

Esse procedimento foi feito para todos os grupos de consultas selecionados (com 1, 2, 3 e 4 palavras-chave). Após a definição do gabarito, com os critérios ditos acima, os documentos pertencentes a este são ordenados pela similaridade textual através da função cosseno [Zobel e Moffat 2006]. Esse procedimento foi feito assumindo que, dentro dos documentos relevantes, o mais similar textualmente atende melhor a uma necessidade de informação do usuário.

3.1.3 Inclusão da Localização Espacial

Após a criação do gabarito é possível estabelecer critérios de inclusão da localização espacial (latitude e longitude) em cada documento da coleção Reuters-21578. Esses critérios serão tratados nesta Seção.

Para que os documentos relevantes sejam avaliados quanto à distância, é necessário possuir distâncias Euclidianas diferentes em relação ao local da consulta. Por exemplo, supondo que um gabarito possua dois documentos {D1, D2}; para que esses documentos sejam avaliados quanto à proximidade espacial, a distância Euclidiana do documento D1, em relação ao local da consulta, deve ser diferente da distância Euclidiana do documento D2, em relação ao local da consulta.

Uma solução é incluir cada documento relevante em um raio diferente um do outro em relação ao local da consulta, utilizando intervalos de acesso. O objetivo dessa abordagem é distribuir os documentos relevantes de acordo com a distância e verificar o comportamento da Consulta Espaço-Textual na recuperação destes. Para ilustrar essa solução, a Figura 3.4 apresenta um conjunto de 5 documentos espaço-textuais {D1, D2, D3, D4, D5 } Os documentos D1 e D2 fazem parte do gabarito e estão espacialmente separados em relação ao local da consulta q, cada um utilizando um intervalo de acesso diferente. O documento D2 está posicionado no intervalo de acesso de [101:200] em relação ao local da consulta, pois 0 é o ponto da consulta. O documento D1 está posicionado no intervalo de acesso de [1, 200]. Assim, é estabelecido que os documentos do gabarito ficam separados um do outro por uma determinada distância. O tamanho dessa distância é definida através de um intervalo pré-estabelecido. No exemplo da Figura 3.4, esse intervalo é de [1:100] para o documento D1 e de [101, 200] para o documento D2, aumentando a distância do documento, em relação ao local da consulta, em uma proporção de 100.

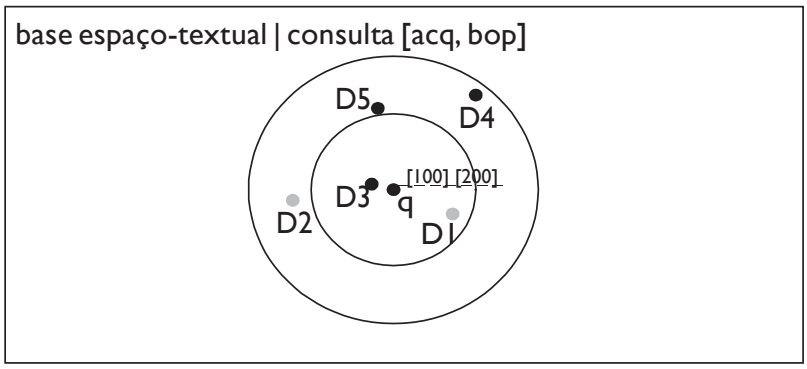


Figura 3.4: Base espaço-textual das palavras-chave “acq” e “bop”.

Os outros documentos, que não fazem parte do gabarito, ficam distribuídos aleatoriamente entre os documentos relevantes, como mostrado na Figura 3.4. Isso é feito para garantir que os demais documentos estejam no mesmo raio de acesso que os itens dentro do gabarito, garantindo que existam tanto documentos relevantes quanto os demais documentos nos arredores do ponto da consulta e com similar distribuição espacial.

Cada consulta selecionada na coleção Reuters-21578 gera uma base espaço-textual, pois os documentos pertencentes ao gabarito precisam estar espacialmente separados em relação ao local da consulta, para serem avaliados quanto à proximidade espacial. Por exemplo, A Figura 3.5 apresenta duas consultas selecionadas da coleção que possui os documentos {D1, D2, D3, D4, D5}. Para a primeira consulta, cujos os termos são “acq” e “bop”, os documentos D1 e D2 fazem parte do gabarito. Eles são são selecionados aleatoriamente, pois nem sempre o mais próximo é o mais similar textualmente em um cenário real. O documento D2 foi selecionado primeiro e ficou posicionado no intervalo de acesso em relação ao ponto da consulta de [1:100], o documento D1 foi selecionado depois e ficou posicionado no intervalo de acesso de [101:200]. Os outros documentos {D3, D4, D5} são distribuídos nos arredores dos documentos relevantes.

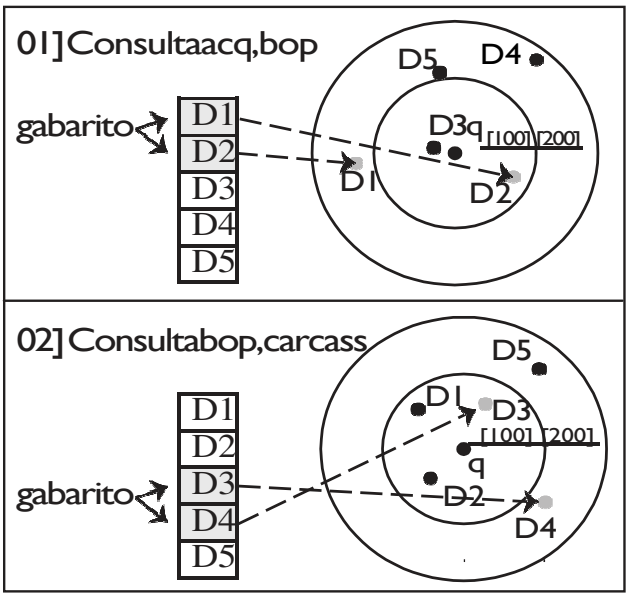


Figura 3.5: Criação de bases espaço-textuais a partir de uma coleção textual.

A segunda consulta apresentada na Figura 3.5, cujos os termos são "bop" e "carcass", os documentos D3 e D4 fazem parte do gabarito, são analisados textualmente com a função cosseno [Zobel e Moffat 2006] e são selecionados aleatoriamente. O documento D3 foi selecionado primeiro e ficou posicionado no intervalo de acesso de [1:100], o documento D4 foi selecionado depois e ficou posicionado no intervalo de acesso [101:200]. Os outros documentos {D1, D2, D5} são distribuídos nos arredores dos documentos relevantes. Com este entendimento, as consultas, ilustradas na Figura 3.5, formam duas bases espaço-textuais, uma para a consulta com as palavras-chave "acq" e "bop" e outra base para a consulta com as palavras-chave "bop" e "carcass".

A representação matemática de como é feita a inclusão da informação espacial, pode ser vista na Figura 3.6. Utiliza-se a Equação 3.1 (distância Euclidiana) para se obter a distância Euclidiana do ponto aleatório (p_a) até à localização da consulta (p_c) com base nas coordenadas cartesianas do ponto aleatório (x_3, y_3) e do ponto da consulta (x_1, y_1), como mostra a Figura 3.6.

$$distância(p.l, q.l) = \sqrt{(p.l_x - q.l_x)^2 + (p.l_y - q.l_y)^2} \tag{3.1}$$

onde $p.l$ é a localização do documento p e $q.l$ é a localização da consulta q. $\{p.l_x, p.l_y\}$ e $\{q.l_x, q.l_y\}$ são as coordenadas (x, y) de cada localização.

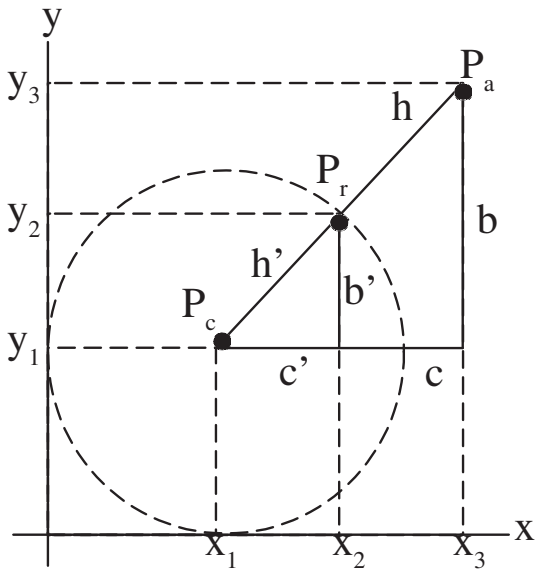


Figura 3.6: Representação matemática da aplicação das teorias de triângulos para os pontos relevantes, sendo p_a (Ponto aleatório), p_r (Ponto relevante) e p_c (Ponto da consulta).

Se o documento selecionado na base textual pertencer ao gabarito, ou seja, esse documento é relevante para a consulta, então, utiliza-se o teorema de Tales, semelhança de triângulos, para estabelecer uma relação matemática entre as localizações dos pontos aleatório (p_a) (x_3, y_3) e o relevante (p_r) (x_2, y_2), para que o ponto relevante (p_r) fique no raio h^j, que é a distância relacionada ao intervalo de acesso. A Equação 3.2 e a Equação 3.3 mostram como o cálculo é feito em relação a Figura 3.6.

$$\frac{h^j}{h} = \frac{b^j}{b} = \frac{c^j}{c} \tag{3.2}$$

$$\frac{h^j}{h} = \frac{x_2 - x_1}{x_3 - x_1} = \frac{y_2 - y_1}{y_3 - y_1} \tag{3.3}$$

A Figura 3.7 apresenta oito documentos, os documentos D1, D2 e D3 fazem parte do gabarito. O documento D3 é selecionado aleatoriamente e é gerada uma localização aleatória, como pode ser visto no quadro 2 da figura. Como o documento D3 faz parte do gabarito, usa-se o teorema de Tales (Equações 3.2 e 3.3) para deslocar a sua posição para o intervalo de acesso de $[1, 100]$ metros. Esse procedimento é feito para todos os documentos pertencentes ao gabarito (documentos D1, D2 e D3), posicionando-os em seus respectivos intervalos de acesso em relação ao ponto da consulta. O documento D2 ficou posicionado em um intervalo de acesso de $[101, 200]$ em relação ao ponto da consulta e o documento D1 ficou posicionado em um intervalo de acesso $[201, 300]$ em relação ao ponto da consulta. Isso significa

que os documentos relevantes da Figura 3.7 ficam posicionados em intervalos de acesso cujo comprimento aumenta em 100 metros. Em seguida, os documentos que não fazem parte do gabarito, ou seja, os demais documentos são distribuídos aleatoriamente nos arredores dos documentos relevantes.

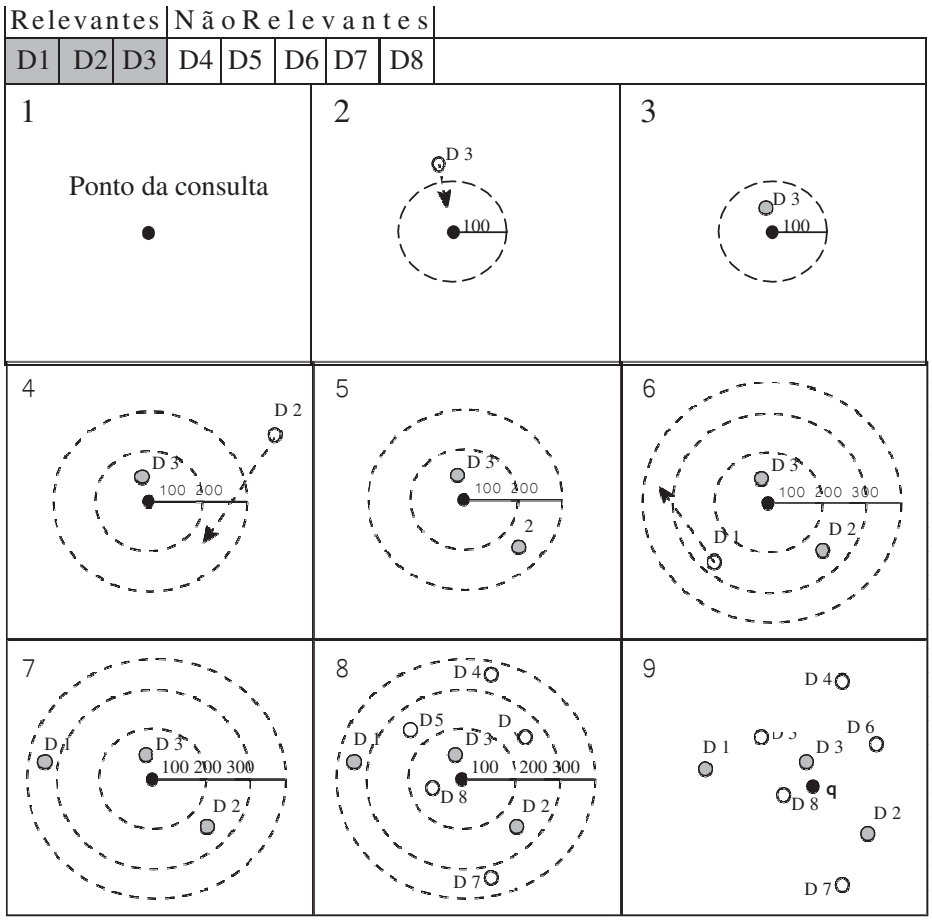


Figura 3.7: Coleção espaço-textual adaptada de uma coleção tradicional.

3.2 Fase de Experimentação

Na fase experimentação, o objetivo é avaliar o desempenho das Equações 2.17 (EQA) e 2.18 (EQB). Para isto, criou-se um conjunto de dados (Seção 3.2.1), definiu-se um conjunto de métricas de avaliação (Seção 3.2.2) e avaliou-se o impacto destas métricas ao variar o número de palavras-chave na consulta, o balanceamento do peso entre similaridade textual e proximidade espacial (α) e o número de objetos esperados como resposta (k).

3.2.1 Conjunto de Dados

Foram criadas 748 coleções espaço-textuais relacionadas às subcategorias da coleção *Reuters*-21578, como mostra a Tabela 3.1. Todas as coleções geradas possuem acima de 5 documentos relevantes em seus respectivos conjuntos de julgamentos de relevância.

Tabela 3.1: Coleções espaço-textuais agrupadas por nº palavras-chave.

Número de palavras-chave	Número de coleções
1	88
2	398
3	190
4	72

As colecões relacionadas às consultas com 1 palavra-chave possuem os maiores números de documentos relevantes em seus gabaritos. Ao passo que, as coleções com 4 palavras-chave possuem os menores números de documentos relevantes em seus respectivos gabaritos. Essa informacão é importante, pois influencia nos resultados, como será visto nos experimentos.

3.2.2 Métricas

As métricas utilizadas na avaliação qualitativa dos resultados são: Precisão, *Average Spatial Similarity* ASS e F_1. A Precisão mede a relevância dos resultados, ASS avalia a distância, ou seja, quanto mais perto os documentos estiverem ao local da consulta, melhor. A métrica F_1 harmoniza em um só valor a métrica Precisão e ASS.

Average Spatial Similarity (ASS) é uma métrica definida nesta pesquisa com o objetivo de avaliar a qualidade dos resultados obtidos nos testes no que se refere à distância. Ela é a média normalizada das distâncias dos documentos relevantes recuperados em relação à localização da consulta. Essa métrica é representada pela Equação 3.4.

$$ASS = 1 - \frac{|AD - min|}{max - min} \tag{3.4}$$

onde o valor de AD (Average Distance) é a média das distâncias entre as localizações dos documentos relevantes recuperados em relação ao local da consulta. A intensão é verificar a qualidade dos resultados quanto à distância dos documentos relevantes recuperados, ou seja, quanto menor for essa média, melhor é a eficácia da Consulta Espaço-Textual Top-k; *min* e *max* são a distância mínima e máxima entre dois documentos da coleção respectivamente.

Nos testes, foram aplicadas as métricas Precisão, ASS e a métrica $F1$, representada pela Equação 3.5, utilizando a métrica Precisão e a métrica ASS, na harmonização, ao invés da utilização da Revocação. Essa métrica é a divisão do número de documentos relevantes recuperados pelo número de documentos do gabarito. No caso deste trabalho, os valores de k não ultrapassaram o número 5. Além disso, uma boa parte das coleções possuem um grande número de documentos nos gabaritos, o que gera sempre um baixo valor de Revocacão, que pode conduzir a falsas conclusões quando essas bases são comparadas às coleções cujo o número de documentos no gabarito é pequeno. Portanto, para evitar tal viés, decidiu-se substituir essa métrica pela métrica ASS, que demonstra melhor a realidade das coleções.

$$F1 = \frac{2 \times Precisão \times ASS}{Precisão + ASS} \tag{3.5}$$

Para exemplificar a aplicação da métrica $F1$, suponha uma coleção de referência espaço-textual formada pelos documentos $\{D1, D2, D3, D4, D5\}$, o gabarito dessa base em relação a uma consulta q seja formado pelos documentos $\{D1, D3, D4\}$ e a distância entre esses documentos e a consulta seja $\{120, 210, 80, 140, 230\}$ (valores em metros). Vale ressaltar que na Figura 3.8, a menor distância entre dois pontos é de $min = 70$ (obtida entre os documentos D3 e D4) e a maior distância entre dois pontos é de $max = 500m$ (obtida entre os documentos D2 e D5).

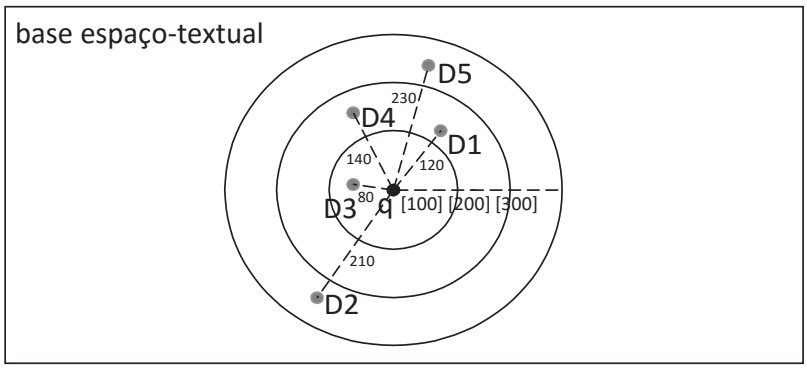


Figura 3.8: Objetos espaço-textuais com distâncias em relação à consulta q.

A consulta q retornou os seguintes documentos $\{D5, D4, D1\}$, nesta ordem. A aplicação detalhada da métrica é:

- O resultado da $Precisão = \frac{2}{3} \cong 0.6$, pois os documentos D1 e D4 foram retornados e fazem parte do gabarito;
- A média das distâncias entre os documentos relevantes recuperados (pertencentes ao gabarito) foi $AD = \frac{(140 + 120)}{2} = 130$;
- O resultado da métrica *Average Spatial Similarity* foi $ASS = 1 - \frac{130 - 70}{500 - 70} \cong 0.87$;

- $Medida\text{-}F_1 = \frac{2 \times 0.6 \times 0.87}{0.6 + 0.87} = 0.71.$

O resultado da métrica F_1, apresentado no exemplo, significa que a relevância espaço-textual dos documentos recuperados pela consulta em relação à coleção do exemplo, foi de 71%.

3.3 Fase de Análise dos Resultados

Esta fase visa entender o que os resultados recuperados dizem sobre o sistema. Dessa forma, através da aplicação das métricas é possível analisar a qualidade dos resultados da Consulta Espaço-Textual Top-k utilizando as duas funções de ranqueamento a fim de responder às questões de pesquisa. Dessa forma, a análise pode ser feita através de gráficos como Histograma de Precisão, como ilustra a Figura 3.9, ou gráficos de Mapa de Calor, como o da Figura 3.10.

O Histograma de Precisão é geralmente utilizado para comparar duas funções de ranqueamento [Baeza-Yates e Ribeiro-Neto 2013]. Mediante desse tipo de gráfico é possível identificar qual a equação obteve o melhor desempenho qualitativo. Por exemplo, na Figura 3.9, os valores positivos são da Equação EQ1 e os valores negativos sao da Equação EQ2. É possível verificar, neste gráfico, que a Equação EQ1 foi mais eficaz (em relação à Precisão) que a Equação EQ2 na maioria das 10 consultas demonstradas na figura. Os valores obtidos para as consultas 1, 3 e 5 são iguais a zero, pois os resultados encontrados utilizando as duas funções de ranqueamento foram equivalentes.

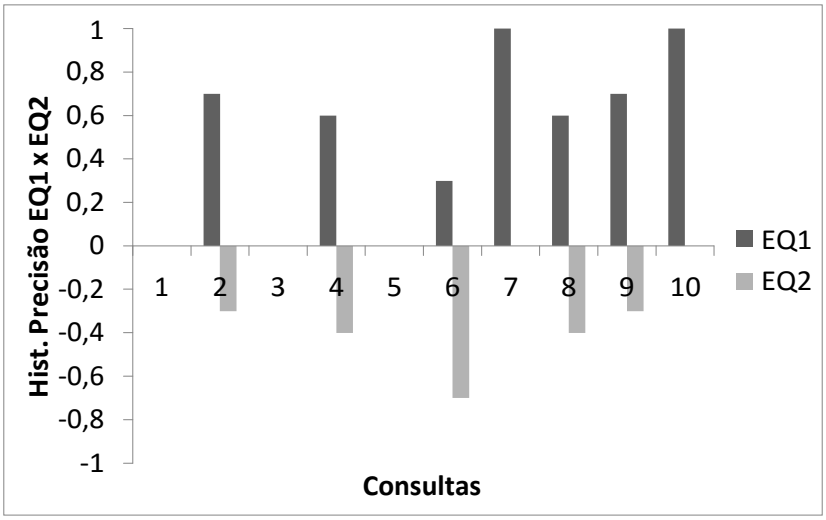


Figura 3.9: Exemplo de Histograma de Precisão.

O gráfico de Mapa de Calor fornece ao avaliador uma ampla visão do comportamento de um determinado resultado [Calumby 2016]. Por exemplo, no gráfico da Figura 3.10 o eixo x representa valores de α da função de ranqueamento de uma Consulta Espaço-Textual Top-k, o eixo y representa valores de k e a barra vertical, posicionada na lateral direita, um *range* de 0.1 a 0.6 da Média da métrica F_1. Com as cores do Mapa de Calor, é possível perceber facilmente a eficácia de um resultado.

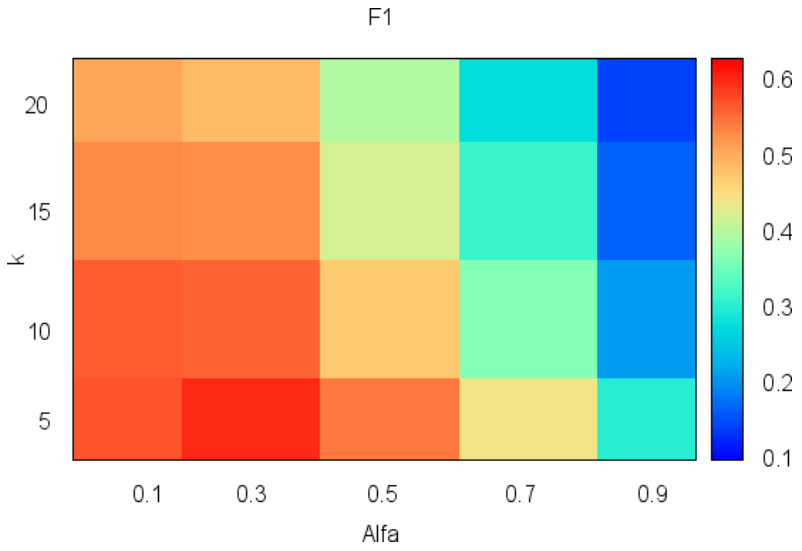


Figura 3.10: Exemplo de Mapa de Calor.

Outros tipos gráficos podem ser utilizados com o intuito de melhorar as análises dos resultados recuperados por uma ou mais consultas, como por exemplo: gráficos de barras, de linhas com marcadores, etc.

Capítulo 4

Avaliação Qualitativa da Consulta Espaço-Textual Top-k

"Não é na ciência que está a felicidade,
mas na aquisição da ciência."
– Edgar Allan Poe

Este capítulo aponta os resultados obtidos no estudo qualitativo da Consulta Espaço-Textual Top-k. A Seção 4.1 apresenta os testes realizados utilizando as métricas Precisão, ASS e F_1, variando o valor de α. A Seção 4.2 exibe os testes realizados utilizando as métricas selecionadas, variando o valor de k. A Seção 4.3 ilustra os resultados variando o número de palavras-chave nas métricas selecionadas. Por fim, a Seção 4.4 apresenta os mapas de calor.

A Tabela 4.1 apresenta os valores selecionados para cada parâmetro. Devido ao grande número de experimentos foram selecionados valores padrões, representados em negrito na Tabela 4.1, para apresentação dos resultados, de maneira estatisticamente significativa.

Nos experimentos realizados, foram escolhidos valores de α equivalentes para cada Equação, pois cada uma tem a sua particularidade. A Equacão 2.17 [Cong et al. 2009] (EQA) multiplica o valor de α pela distância normalizada e a Equação 2.18 [Rocha-Junior e Nørvåg 2012] (EQB) faz a multiplicação do valor de α diretamente pela distância Euclidiana, sem normalização.

Tabela 4.1: Tabela dos Parâmetros.

Parâmetro	**Valores Utilizados**
k	1, 2, **3**, 4, 5
Número de Palavras-chave	1, **2**, 3, 4
α EQA	0.01, 0.03, **0.05**, 0.07, 0.09
α EQB	0.001, 0.003, **0.005**, 0.007, 0.009

Os experimentos iniciais, para a Equação [Cong et al. 2009] (EQA), foram realizados com os seguintes valores de α: 0.1, 0.3, 0.5, 0.7, 0.9. A Tabela 4.2 apresenta a precisão utilizada pela equação EQA para estes valores de α. Como os resultados reagiram negativamente, procurou-se um conjunto de valores cuja Precisão proporcionassem melhores resultadores e optou-se pelos valores apresentados na Tabela 4.1.

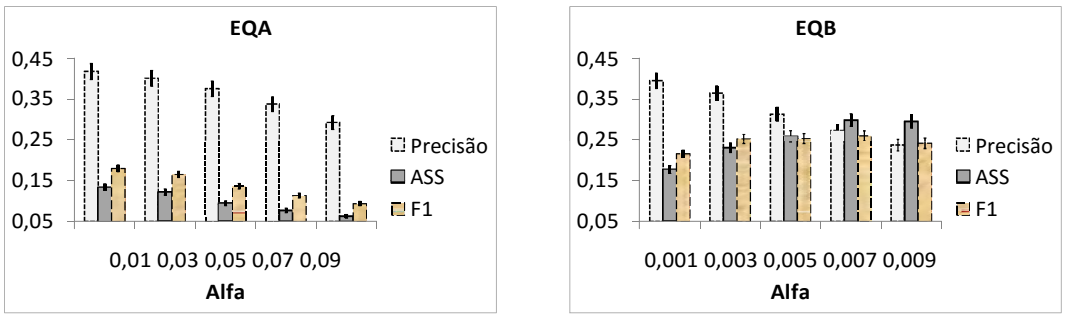
Tabela 4.2: Valores de Precisão Realizados nos Experimentos Iniciais para a Equação EQA.

Alfa	0.1	0.3	0.5	0.7	0.9
EQA/P@k	0,249	0,013	0,008	0,008	0,014

A primeira constatação que pode ser feita é que o valor de $\alpha = 0.5$, que aparentemente indica peso igual para similaridade textual e proximidade textual, não funcionou para balancear essas duas medidas, porque elas são de naturezas diferentes, requerendo valores mais específicos. A mesma abordagem foi adotada para selecionar os valores de α para a Equação EQB (Tabela 4.1).

4.1 Variando α

Esta seção apresenta os resultados qualitativos, variando os valores de α, utilizando as duas funções de ranqueamento existentes para a Consulta Espaço-Textual Top-k. O objetivo dessa variação é verificar o impacto qualitativo, dos valores de α selecionados para cada equação, na recuperação de documentos. A Figura 4.1 ilustra o resultado encontrado variando o valor de α, com o valor de $k = 3$ e o número de palavras-chave $= 2$ utilizando as duas funções de ranqueamento existentes para a Consulta Espaço-Textual Top-k.



(a) EQA. (b) EQB.

Figura 4.1: Variando o valor de α.

Com a utilização dos valores de α selecionados para as duas equações (EQA e EQB), é possível perceber na figura que os melhores valores de Precisão ocorreram quando os valores de α aplicam uma maior importância à similaridade textual (os valores 0.01 e 0.03 para a equação EQA e os valores 0.001 e 0.003 para a equação EQB). À medida que o valor de α aumenta, a Precisão diminui, pois menos documentos relevantes são recuperados. Para esta métrica, os melhores valores de Precisão foram da Equação EQA, pois ocorreram uma maior recuperação de documentos relevantes do que com a utilização da Equação EQB.

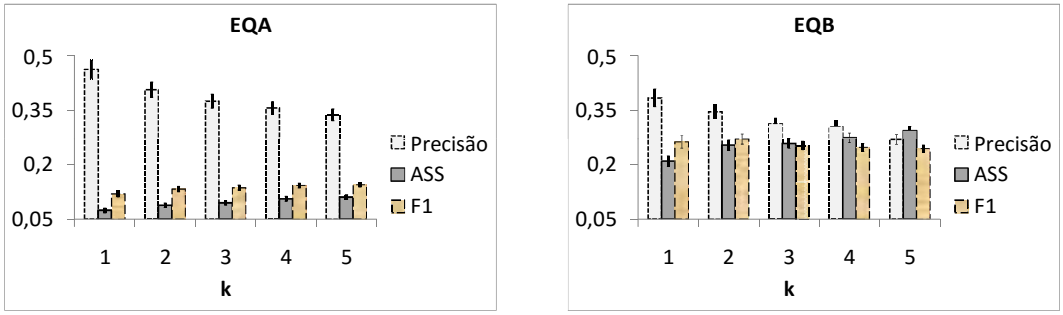
Os resultados obtidos com a aplicação da métrica *Average Spatial Similarity* (ASS) utilizando a Equação EQA foram inesperados, pois à medida que o valor de α aumenta espera-se que a Consulta Espaço-Textual Top-k recupere mais documentos relevantes próximos ao local da consulta, no entanto, como mostra a Figura 4.1(a), na barra relacionada à métrica ASS, o aumento do valor de α fez com que as consultas recuperassem menos documentos relevantes próximos ao local da consulta.

No que se refere aos resultados obtidos com a equação EQB, os documentos relevantes recuperados estavam mais próximos da localização da consulta, como mostra a Figura 4.1(b) na barra relacionada à métrica ASS. Ou seja, o aumento do valor de α proporcionou uma melhor recuperação espacial, pois as consultas trouxeram mais documentos relevantes próximos ao local da consulta.

Os resultados obtidos com a aplicação da métrica F_1, que harmoniza os valores de Precisão e ASS, indicaram que a Equação EQB foi mais eficaz, pois além de recuperar documentos relevantes, esses documentos estavam mais próximos ao local da consulta, como mostra a Figura 4.1(b).

4.2 Variando k

Esta seção apresenta os resultados dos testes realizados variando o valor de k, utilizando o valor de $\alpha = 0.05$ para a equação EQA, o valor de $\alpha = 0.005$ para a equação EQB e o número de palavras chave = 2. O intuito desta variação é verificar o impacto qualitativo da Consulta Espaço-Textual Top-k quando mais documentos são recuperados. A Figura 4.2 ilustra os resultados obtidos variando o valor de k utilizando as duas equações (EQA e EQB).



(a) EQA. (b) EQB.

Figura 4.2: Variando o valor de k.

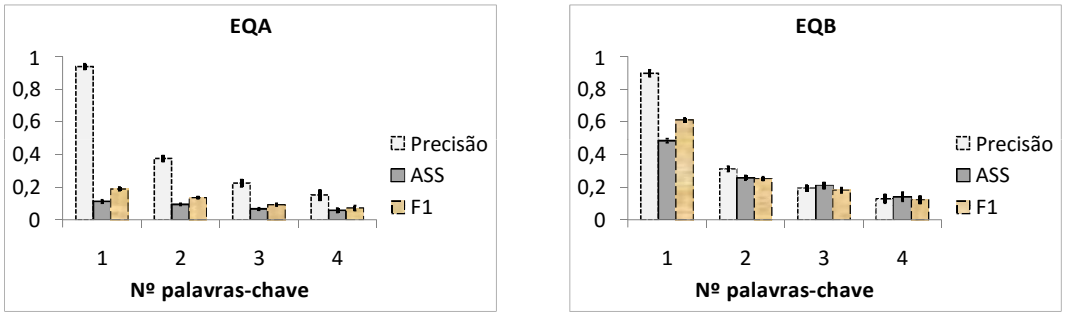
Como mostra a Figura 4.2, os valores de Precisão foram melhores com a utilização da equação EQA, isto é, mais documentos relevantes foram recuperados com esta equação. Além disso, o aumento do valor de k proporcionou uma diminuição da Precisão utilizando as duas equações, como mostra as Figuras 4.2(a) 4.2(b). Esse comportamento ocorreu porque menos documentos relevantes foram recuperados.

Em relação à qualidade espacial dos documentos relevantes retornados pelas consultas, à proporção que o valor de k aumentava, mais documentos relevantes próximos ao local da consulta foram recuperados no emprego das duas equações. A qualidade espacial dos documentos relevantes utilizando a equação EQB foi superior à equação EQA, como mostra a Figura 4.2(b).

Comparando a qualidade de recuperacão espaco-textual das duas equações (EQA e EQB) através da métrica F_1, a equação EQB indicou uma maior eficácia. Assim sendo, é possível inferir que com os valores dos parâmetros utilizados, a equacão EQB tem uma capacidade de recuperação espaço-textual maior do que a equação EQA, pois para uma Consulta Espaço-Textual espera-se que os documentos relevantes retornados encontrem-se próximos à localização da consulta.

4.3 Variando Nº Palavras-Chave

Esta seção visa apresentar os resultados em uma perspectiva diferente dos demais, variando o número de palavras-chave. Neste contexto, os valores padrões dos parâmetros selecionados foram: $k = 3$, $\alpha = 0.05$ para a equação EQA e $\alpha = 0.005$ para a equação EQB. A Figura 4.3 apresenta o resultado obtido com a variação do número de palavras-chave utilizando as duas equacões existentes para a Consulta Espaço-Textual Top-k.



(a) EQA. (b) EQB.

Figura 4.3: Variando o valor do Nº de Palavras-Chave.

As consultas com uma palavra-chave tiveram melhores resultados qualitativos que as demais consultas, como pode ser visto na Figura 4.3. Isto aconteceu pois a quantidade de documentos relevantes relacionados às consultas com uma palavra-chave é maior que as outras consultas, aumentando a probabilidade de um documento relevante ser recuperado. Por este motivo, ocorreu uma diminuição do valor de Precisão nos resultados com o aumento do valor do número de palavras-chave.

A equação que proporcionou uma melhor qualidade espacial foi a EQB, como pode ser visto na Figura 4.3(b) nas barras relacionadas à métrica ASS. Isso significa que independente do número de palavras-chave utilizado em uma consulta, a equação EQB recupera mais documentos relevantes próximos ao local da consulta do que a equação EQA.

A equação EQB proporcionou melhores resultados qualitativos para uma Consulta Espaço-Textual, como pode ser verificado na Figura 4.3(b) nas barras relacionadas à métrica F_1. Portanto, diante dos valores dos parâmetros selecionados, a equacão EQB foi mais eficaz que a equação EQA.

4.4 Correlação entre as Métricas

Esta seção apresenta mapas de calor para verificação da correlação entre as métricas utilizadas nesta pesquisa (Precisão, ASS e F_1). A Figura 4.4 ilustra mapas de calor relacionados às consultas com 3 palavras-chave, variando o valor de k e α. O eixo x se refere aos valores de α, o eixo y se refere aos valores de k.

O intuito da utilização desses mapas não é comparar as duas funções de ranqueamento, pois as escalas apresentadas são diferentes. São usadas escalas apropriadas em cada mapa, para destacar o comportamento dos testes realizados, considerando os efeitos causados pela variação do valor de α e k em cada equação. É possível perceber na Figura 4.4 um comportamento diferente entre as duas equações para todas as métricas aplicadas.

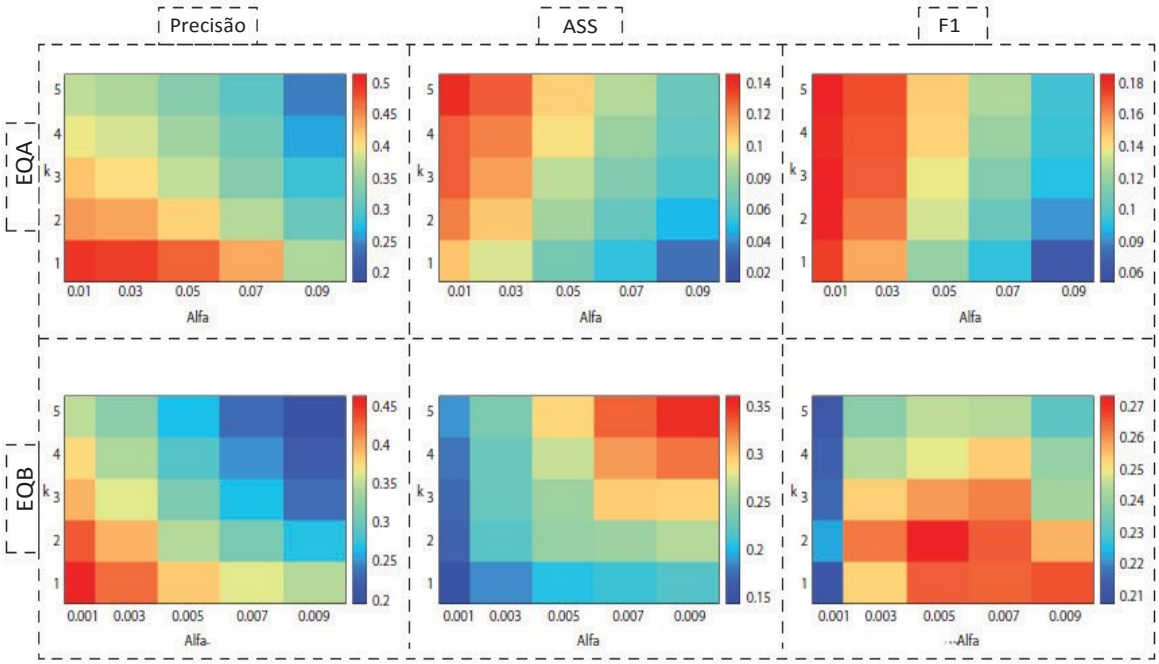


Figura 4.4: Mapa de Calor correlacionado às métricas.

Em todos os mapas relacionados à Equação EQA, os melhores resultados ocorreram com o valor de $\alpha = 0.01$, que atribui uma maior importância à similaridade textual. Isso indica que a capacidade de recuperar documentos relevantes, utilizando a Equação EQA, aumenta com a diminuição do valor de α, nas consultas selecionadas. Logo, esta consulta favorece maior preferência à similaridade textual.

Analisando a Equação EQA para a métrica ASS, os resultados menos significativos ocorreram com o valor de $\alpha = 0.09$, indicando que o aumento da importância da proximidade espacial não garantiu documentos mais próximos ao local da consulta

nos resultados. Além disso, os melhores resultados ocorreram com o valor de $k = 5$ e $\alpha = 0.01$, demonstrando que o aumento de k aumenta o número de documentos relevantes recuperados.

No que se refere à Equação EQB, analisando o mapa relacionado à métrica ASS, com o aumento dos valores de α, mais documentos relevantes próximos ao local da consulta foram recuperados. Os resultados que indicaram uma alta capacidade de recuperar documentos próximos ao local da consulta ocorreram com os valores de $\alpha = 0.009$ e $k = 5$. A respeito da métrica Precisão para esta equação, os melhores resultados ocorreram com os valores de $\alpha = 0.001$ e $k = 1$. Além disso, com a aplicação da métrica F_1, os melhores resultados ocorreram com o $\alpha = 0.005$, demonstrando uma harmonia entre as duas métricas (Precisão e ASS).

Capítulo 5 Considerações

Finais

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes."

– Marthin Luther King

Este trabalho avaliou qualitativamente a Consulta Espaço-Textual Top-k. Esta consulta retorna os k melhores resultados baseados na proximidade espacial e na similaridade textual dos documentos retornados em relação ao local da consulta. Ela utiliza três parâmetros: k (número de documentos retornados pela consulta), palavras-chave da consulta e a localização da consulta.

Para tanto, a consulta é executada usando uma função de ranqueamento, que no caso deste trabalho foram utilizadas as duas funções existentes ([Cong et al. 2009, Rocha-Junior e Nørvåg 2012]. Ambas consideram a variável α que exerce uma influência na relação entre a proximidade espacial e a similaridade textual. A depender do valor desta variável o resultado da consulta pode ser impactado qualitativamente, ou seja, os resultados podem não atender a uma necessidade de informação do usuário, diminuindo a eficácia da consulta.

Com o objetivo de avaliar o quão sensível é a variação dos valores dos parâmetros envolvidos neste sistema, foram desenvolvidas coleções de teste espaco-textuais a partir de coleções tradicionais, como a Reuters-21578, utilizando o paradigma de *Cranfield*, que estabelece que uma coleção de teste deve possuir um conjunto de documentos, um conjunto de consultas e um conjunto de julgamentos de relevância ou gabaritos. Para isto, foi desenvolvida uma proposta de metodologia de avaliação que possui as etapas de preparação, experimentacão e análise.

Na fase de preparação foram criadas 748 coleções (*download*:

https://goo.gl/82zuVK) de teste espaço-textuais agrupadas por número de palavras chave (uma, duas, três e quatro). Além disso, foi necessário incluir uma informação espacial (x, y) em cada documento da coleção Reuters-21578, estabelecendo que os documentos relevantes para uma determinada consulta ficam separados um do outro através de um intervalo de acesso, para que a proximidade espacial seja avaliada, pois quanto mais perto um documento está em relação à localização da consulta, mais relevante espacialmente ele é. Os documentos não relevantes ficam distribuídos entre os relevantes, isso é feito para garantir que os itens não relevantes estejam no mesmo raio de acesso que os itens dentro do gabarito, garantindo que existam tanto documentos relevantes quanto documentos não relevantes nos arredores do ponto da consulta e com similar distribuição espacial.

Na fase de experimentação foram utilizadas as métricas Precisão, definida a métrica *Average Spatial Similarity* e $F1$. Precisão é o cálculo da fração do número de documentos relevantes recuperados sobre os documentos recuperados, a métrica *Average Spatial Similarity* é a média dos documentos relevantes em relação ao local da consulta e $F1$ estabelece uma média harmônica entre a Precisão e a ASS. Por fim, a fase de análise é a fase para avaliar os resultados obtidos com a aplicação das métricas utilizadas, ilustrando os resultados através de gráficos como o Histograma de Precisão, os mapas de calor e outros gráficos com a finalidade de verificar o comportamento da Consulta Espaco-Textual Top-k utilizando as duas funções de ranqueamento.

Abaixo são apresentadas as questões de pesquisa com as suas respectivas respostas obtidas através deste trabalho.

Quais métricas podem ser utilizadas para avaliar a consulta espaço-textual de forma qualitativa?

As métricas escolhidas para esta pesquisa foram:

- **Precisão**. Utilizada para avaliar a relevância dos documentos recuperados pela consulta, ou seja, quanto mais documentos pertencentes ao gabarito estiverem em um resultado, mais eficaz é a consulta;
- ***Average Spatial Similarity* (ASS)**. Essa métrica foi definida neste trabalho com o objetivo de se avaliar a proximidade es- pacial dos documentos relevantes recuperados pela consulta. Ela foi fundamental para avaliar a Consulta Espaco-Textual Top-$k$no que se refere à capacidade das funções de ranqueamento (Equações EQA e EQB) de recuperar documentos relevantes próximos ao local da consulta. Deste modo, quanto maior for o valor dessa métrica, maior será essa capacidade.
- **F_1**. A métrica F_1 funciona para harmonizar duas métricas, geralmente Precisão e Revocação, em um só valor. Para este trabalho,

a F_1 foi utilizada com a Precisão e a ASS. Quanto maior for esse valor, maior é a qualidade dos resultados da consulta em relação a essas duas medidas.

Qual a qualidade dos resultados, quando comparada com métricas que avaliam a relevância dos resultados retornados, métricas que avaliam a distância e métricas que usam harmonicamente as duas medidas?

A seleção das métricas selecionadas teve como objetivo entender a qualidade dos resultados recuperados pela consulta sob três aspectos: a capacidade de retornar documentos relevantes (Precisão); a qualidade dos resultados quanto à proximidade espacial (ASS) e; a qualidade dos resultados quando as métricas Precisão e ASS são harmonizadas (F_1).

Com os resultados obtidos com a métrica Precisão foi possível identificar o comportamento das funções de ranqueamento no que se refere à recuperação de documentos relevantes. De todos os tipos de gráficos utilizados para esta métrica, o Histograma de Precisão foi fundamental para verificar o desempenho qualitativo das duas Equações (EQA e EQB), pois ilustrou, de maneira clara, a Precisão em cada consulta.

A qualidade dos resultados quanto à proximidade especial foi demonstrada com a aplicação da métrica *Average Spatial Similarity*. Os gráfi- cos apresentados no Cap´ıtulo 4 permitiram observar o comportamento das duas equações, identificando qual equação obteve melhor desempenho qualitativo neste quesito.

A métrica F_1 foi importante para analisar a qualidade de recuperação das duas equações quando as métricas Precisão e ASS são harmonizadas. Foi possível, portanto, comparar as equações e identificar qual delas obteve um desempenho melhor, recuperando documentos relevantes e mais próximos ao local da consulta para o usuário.

Qual função de ranqueamento melhor atende a uma necessidade de informação do usuário?

Os resultados obtidos com a métrica Precisão, utilizada para avaliar a relevância dos resultados retornados, mostraram que a Equacão EQA foi mais eficaz que a Equação EQB. Com relação aos resultados obtidos com a métrica ASS, utilizada para avaliar a qualidade dos resultados quanto à proximidade espacial do documento recuperado ao local da consulta, a Equação EQB foi mais eficaz que a Equação EQA. Por fim, no que se refere à harmonização dessas duas métricas (F_1), a Equacão EQB obteve melhores resultados que a Equação EQA.

Dessa forma, se a necessidade de informacão do usuário perpassa pela recuperação de documentos relevantes independentemente da distância deles em relação ao local da consulta, a melhor equacão foi a EQA.

> Em contrapartida, se esta necessidade for atendida com a recuperação de documentos relevantes mais próximos ao usuário, a melhor equação foi a EQB.

Portanto, para executar a metodologia estabelecida nesta pesquisa, no intuito de avaliar qualitativamente a Consulta Espaco-Textual Top-k, foi necessária a criação de uma coleção de teste espaço-textual com um conjunto de documentos espaço-textuais, um conjunto de consultas e um conjunto de julgamentos de relevância. Foram exigidas também métricas que avaliaram a relevância dos resultados retornados pela consulta, como a métrica de Precisão (que verifica a relevância o resultado), a métrica *Average Spatial Similarity* (que avalia a proximidade espacial, estabelecendo que quanto mais perto o documento relevante estiver em relação à localização da consulta, mais relevante espacialmente esse documento será para o usuário) e a métrica F_1 (que harmoniza essas duas medidas em um só valor). Por fim, a pesquisa proporcionou uma importante base aos interessados em avaliar um sistema de recuperação espaço-textual.

5.1 Pesquisas Futuras

Abaixo são apresentados algumas possibilidades de pesquisas futuras a partir do que foi estudado neste trabalho:

- **Otimização da variável** α. Otimizar o valor de α para encontrar valores que proporcionem melhores resultados qualitativos. Pode ser utilizado, neste caso, programação linear ou algoritmos genéticos.
- **Utilização de bases não sintéticas com julgamentos de relevância textual e espacial**. Utilizar coleções de teste cujos julgamentos de relevância estão relacionados às perspectivas textuais e espaciais feitos por um conjunto de pessoas. A ideia é verificar o comportamento das consultas espaco-textuais em um ambiente real.

Referências Bibliográficas

[Baeza-Yates e Ribeiro-Neto 2011]Baeza-Yates, R. e Ribeiro-Neto, B. (2011). *Mo- dern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

[Baeza-Yates e Ribeiro-Neto 2013]Baeza-Yates, R. e Ribeiro-Neto, B. (2013). Modern information retrieval. *ACM Press. New York*.

[Borlund 2003]Borlund, P. (2003). The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):3–8.

[Brin e Page 1998]Brin, S. e Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.

[Calumby 2016]Calumby, R. T. (2016). Diversity-oriented multimodal and interactive information retrieval. *SIGIR Forum*, 50(1):86.

[Cao et al. 2012]Cao, X. C., Cong, G., Jensen, C. S., Qu, Q., Skovsgaard, A., Wu, D., e Yiu, M. L. (2012). Spatial keyword querying. *Er*, 7532(1):16–29.

[Cardoso 2007]Cardoso, N. (2007). Gikip - geoclef pilot for crosslingual geographic information retrieval from wikipedia. http://www.linguateca.pt/GikiP/. [On- line; acessado 04-agosto-2016].

[Cary et al. 2010]Cary, A., Wolfson, O., e Rishe, N. (2010). Efficient and scalable method for processing top-k spatial boolean queries. *SSDBM*, 6187:87–95.

[Chapelle et al. 2012]Chapelle, O., Joachims, T., Radlinski, F., e Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Tran- sactions on Information Systems*, 30(1):1–41.

[Chen et al. 2013]Chen, L., Cong, G., Jensen, C. S., e Wu, D. (2013). Spatial keyword query processing: An experimental evaluation. *Proceedings of the VLDB Endowment*, 6(3):217–228.

[Chen et al. 2006]Chen, Y.-Y., Suel, T., e Markowetz, A. (2006). Efficient query processing in geographic web search engines. *SIGMOD*, pp. 277–288.

[Christoforaki et al. 2011]Christoforaki, M., He, J., Dimopoulos, C., e Marjowetz, A. (2011). Text vs. space: efficient geo-search query processing. *CIKM*, pp. 423–432.

[Cleverdon 1991]Cleverdon, C. W. (1991). The significance of the cranfield tests on index languages. *SIGIR '91 Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12.

[Cong et al. 2009]Cong, G., Jesen, C. S., e Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. *Proceedings of the VLDB Endowment*, 2(1):337–348.

[Cooper 1994]Cooper, W. S. (1994). The formalism of probability theory in IR: a foundation or an encumbrance? In Croft, W. B. e van Rijsbergen, C. J., editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 242–247. Springer.

[Croft e Lafferty 2003]Croft, W. B. e Lafferty, J. (2003). *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA.

[Felipe et al. 1984]Felipe, I. D., Hristidis, V., e Rishe, N. (1984). Keyword search on spatial databases. *ICDE*, pp. 656–665.

[Gey et al. 2005]Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., e Petras, V. (2005). The clef 2005 cros-language geographic information retrieval track overview. *Lecture Notes in Computer Science*, pp. 908–919.

[Göker e Myrhaug 2008]Göker, A. e Myrhaug, H. (2008). Evaluation of a mobile information system in context. *Pergamon Press*, 44(1):39–65.

[Hariharan et al. 2007]Hariharan, R., Hore, B., Li, C., e Mehrota, S. (2007). Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems. *IEEE Computer Society Washington*, pp. 16.

[Hersh et al. 1994]Hersh, W., Buckley, C., Leone, T. J., e Hickam, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. *SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 192–201.

[Hilbert 1981]Hilbert, D. (1981). über die stetige abbildung einer line auf ein flaächenstück. *Mathematische Annalen*, 38(3):459–460.

[Jones 1981]Jones, K. S. (1981). *Information Retrieval Experiment*. Butterworth-Heinemann Newton, MA, USA.

[Jones et al. 2000]Jones, K. S., Walker, S., e Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. In *Information Processing and Management*, pp. 779–840.

[Khodaei et al. 2010]Khodaei, A., Shahabi, C., e Li, C. (2010). Hybrid indexing and seamless ranking of spatial and textual features of web documents. *DEXA*, pp. 450–466.

[Kleinberg 1999]Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

[Lewis 2004]Lewis, D. D. (2004). Reuters-21578 text categorization text collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/. [Online; acessado 19-maio-2016].

[Lugo e Alberto 2004]Lugo, G. e Alberto, G. (2004). *Um modelo de sistemas multiagentes para partilha de conhecimento utilizando redes sociais comunitárias*. PhD thesis, Escola Politécnica da Universidade Estadual de São Paulo, USP, SP. 10.11606/T.3.2004.tde-15112004-190053.

[Manning et al. 2008]Manning, C. D., Raghavan, P., e Sch ütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Meadow et al. 2008]Meadow, C. T., Boyce, B. R., Kraft, D. H., e Barry, C. L. (2008). *Text Information Retrieval Systems*.

[Mitchell 1996]Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. A Bradford Book The MIT Press, Cambridge, Massachusetts.

[Müller et al. 2010]M üller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Jr., C. E. K., e Hersh, W. (2010). Overview of the clef 2009 medical image retrieval track. *Multilingual Information Access Evaluation II. Multimedia Experiments*, 6242(1):72–84.

[MOOEM 1951]MOOEM, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *AMERICAN DOCUMENTATION*, 2:20–32.

[Papadias et al. 2001]Papadias, D., Kalnis, P., Zhang, J., e Tao, Y. (2001). Efficient olap operations in spatial data warehouses. *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases (SSTD)*, pp. 442–459.

[Paramita et al. 2007]Paramita, M. L., Sanderson, M., e Clough, P. (2007). Developing a test collection to support diversity analysis. *SIGIR 2009 Workshop: Redundancy, Diversity, and Interdependent Document Relevance*.

[Reuters-21578 2006]Reuters-21578 (2006). Reuters-21578 test collection 2006. http://www.daviddlewis.com/resources/testcollections/reuters21578/. [Online; acessado 23-setembro-2016].

[Richardson et al. 2006]Richardson, M., Prakash, A., e Brill, E. (2006). Beyond pagerank: Machine learning for static ranking. *Proceedings of WWW*, pp. 707–715.

[Rocha-Junior 2012]Rocha-Junior, J. B. (2012). *Efficient Processing of Preference Queries in Distributed and Spatial Databases*. NTNU, Norwegian University of Science and Technology.

[Rocha-Junior et al. 2011]Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., e Nørvåg, K. (2011). Efficient processing of top-k spatial keyword queries. *Lec- ture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6849 LNCS(1):202–222.

[Rocha-Junior e Nørvåg 2012]Rocha-Junior, J. B. e Nørvåg, K. (2012). *Top-k spatial keyword queries on road networks*. Proceedings of the 15th International Conference on Extending Database Technology - EDBT 12, Norwegian University of Science and Technology.

[Russell e Norvig 2003]Russell, S. J. e Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition.

[Sagan 1994]Sagan, H. (1994). *Space-Filling Curves*. Springer-Verlag, Berlin.

[Salton e Buckley 1988]Salton, G. e Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

[Salton e McGill 1986]Salton, G. e McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw Hill Book Co.

[Sanderson 1994]Sanderson, M. (1994). Reuters test collection. *Glasgow University Computing Science Department*.

[Vaid et al. 2005]Vaid, S., Jones, C. B., Joho, H., e Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. *SSTD*, pp. 218–235.

[Veloso et al. 2008]Veloso, A. A., Almeida, H. M., Goncalves, M. A., e Jr., W. M. (2008). Learning to rank at query-time using association rules. *Intl ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 267–274.

[Wilkinson e Wu 2004]Wilkinson, R. e Wu, M. (2004). Evaluation experiments and experience from the perspective of interactive information retrieval. *the Procee- dings of the Third Workshop on Empirical of Adaptive Systems*, pp. 23–26.

[Witten et al. 1999]Witten, I. H., Moffat, A., e Bell, T. C. (1999). *Managing Gi- gabytes: Compressing and Indexing Documents and Images, Second Edition*. Mor- gan Kaufmann.

[Yan et al. 2009]Yan, H., Ding, S., e Suel, T. (2009). Inverted index compression and query processing with optimized document ordering. *WWW*, pp. 401–410.

[Zhou et al. 2005]Zhou, Y., Xie, X., Wang, C., Gong, Y., e Ma, W.-Y. (2005). Hybrid index structures for location-based web search. *CIKM*, pp. 155–162.

[Zobel e Moffat 1998]Zobel, J. e Moffat, A. (1998). Exploring the similarity space. In *ACM SIGIR Forum*, volume 32, pp. 18–34. ACM.

[Zobel e Moffat 2006]Zobel, J. e Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys (CSUR)*, 38(2).

[Zobel et al. 2011]Zobel, J., Webber, W., Sanderson, M., e Moffat, A. (2011). Principles for robust evaluation infrastructure. *Proceedings of the 2011 workshop on Data infrastructurEs for supporting information retrieval evaluation*, pp. 3–6.