UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Richard Henrique de Souza

QSM: UMA ABORDAGEM PARA RECUPERAR E ORDENAR QUESTIONÁRIOS

Florianópolis
2020

Richard Henrique de Souza

QSM: UMA ABORDAGEM PARA RECUPERAR E ORDENAR QUESTIONÁRIOS

Tese submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de doutor em ciência da computação.
Orientadora: Carina Friedrich Dorneles, Dra.

Florianópolis
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Souza, Richard Henrique de
QSM: UMA ABORDAGEM PARA RECUPERAR E ORDENAR QUESTIONÁRIOS / Richard Henrique de Souza ; orientadora, Carina Friedrich Dorneles, 2020.
80 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2020.

Inclui referências.

1. Ciência da Computação. 2. Banco de dados. 3. Recuperação de Informação. I. Dorneles, Carina Friedrich . II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Ciência da Computação. III. Título.

Richard Henrique de Souza

QSM: UMA ABORDAGEM PARA RECUPERAR E ORDENAR QUESTIONÁRIOS

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Daniel S. Kaster, Dr.
Universidade Estadual de Londrina

Prof. Leandro Krug Wives, Dr.
Universidade Federal do Rio Grande do Sul

Prof. Alexandre Gonçalves, Dr.
Universidade Federal de Santa Catarina

Profa. Vania Bogorny , Dra.
Universidade Federal de Santa Catarina

Prof. Roberto Willrich , Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em ciência da computação.

Documento assinado digitalmente
Vania Bogorny
Data: 28/10/2020 08:09:13-0300
CPF: 684.586.570-15

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
Carina Friedrich Dorneles
Data: 28/10/2020 09:13:30-0300
CPF: 921.834.250-91

Carina Friedrich Dorneles, Dra.
Orientadora

Florianópolis, 2020.

Este trabalho é dedicado à minha esposa e filho.

AGRADECIMENTOS

Agradeço ao pessoal da SeTIC pelo apoio.

“ Se você quer ser bem sucedido, precisa ter dedicação total, buscar seu último limite e dar o melhor de si.”

Ayrton Senna

RESUMO

Questionários são ferramentas úteis para fins de pesquisa e geralmente são usados para coletar informações sobre uma população de interesse, concentrando-se em diferentes intenções. Durante o projeto do questionário, ou reuso de dados obtidos em outras pesquisas, pode ser útil verificar se já existe um questionário com intenção similar que o proposto. Perguntas bem projetadas podem levar os entrevistados a fornecer melhores respostas. No entanto, a busca por questionários de pesquisa não é uma tarefa trivial, pois uma pergunta pode ser estruturada de maneiras diferentes. Neste trabalho, propõe-se uma abordagem para recuperar questionários por meio do cálculo de similaridade entre questionários. A abordagem considera a heterogeneidade das perguntas e fornece um método de classificação baseado nas diferentes possibilidades de consultas. A abordagem considera o questionário em formato de árvore, realiza o cálculo de similaridade de cada folha e em seguida realiza uma média ponderada da similaridade das folhas para obter a similaridade entre questionários. Para determinar a eficácia dessa abordagem, é realizada uma série de experimentos, utilizando revocação, precisão, valor-f, MAP e NDGC, a partir dos quais foi constatado que a abordagem proposta tem uma precisão em torno de 40% melhor do que os modelos tradicionais testados. A revocação obteve um resultado em torno 20% melhor. Nas métricas f-value, MAP e NDCG a abordagem proposta obteve um resultado em torno em 30% melhor.

Palavras-chave: Questionário. Recuperação de informação. Similaridade.

ABSTRACT

Questionnaires are useful tools for research purposes and are generally used for collecting information about a population of interest, by focusing on different intentions. During the questionnaire project, or for sharing data purposes, it may be useful to check if there is already a questionnaire with the same intention as that being carried out. Well-designed questions can induce respondents to provide better answers. However, examining research questionnaires is not a trivial task since a question can be structured in different ways. In this work, we propose a similarity measure to match questionnaires that are characterized by the heterogeneity of their questions and to provide a ranking method based on variations of a given query. The approach considers the questionnaire as a tree, calculates the similarity of each leaf and then performs a weighted average of the similarity of the leaves to obtain the similarity between questionnaires.To determine the effectiveness of this approach, a series of experiments is carried out, using recall, precision, f-value, MAP and NDGC, from which it was found that the proposed approach has an accuracy around 40% better than the models traditionally tested. The recall had a result around 20% better than others. In the f-value, MAP and NDCG metrics the proposed approach achieved a result around 30% better than others.

Keywords: Questionnaire, Information retrieval, Similarity.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

a	Uma opção de resposta (uma alternativa)
A	Conjunto de opções de resposta (Alternativas)
CQA	Community Question Answering
CQAs	Community Question Answering sites
Ctes	Conjunto de TES
Ctss	conjunto de valores de TSS
DCG	Discounted Cumulated Gain
e	Enunciado da pergunta
FAQ	Frequently Asked Questions
F-value	Média harmônica da precisão
IDF	Inverse Document Frequency
L	Linguagem
MAE	Mean Absolute Error
MAP	Mean Average Precision
MSL	Mapeamento Sistemático de Literatura
NDCG@k	Normalized Discounted Cumulated Gain at top k positions
NGD	Normalized Google Distance
P	Precisão
PICOC	Population, Intervention, Comparison, Outcome, Context
Q	Questionário
QSM	Questionnaire Similarity Matching
R	Revocação
T	Função de tokenização
TES	Token Equal Score
TF	Term Frequency
TSS	Token Synonym Score
WLM	Wikipedia-based similarity metric

LISTA DE SÍMBOLOS

$\in$	Pertence
$\rightarrow$	Vetor
$\sum$	Somatória
$\lvert\ \rvert$	Módulo
$\cup$	União
O	Limite assintótico superior
$\leftarrow$	Atribuição
$\cup$	União

SUMÁRIO

1 INTRODUÇÃO

De acordo com Kothari (2004), a pesquisa científica usualmente necessita coletar dados, sendo que a tarefa de coleta de dados começa depois da definição de um problema da pesquisa. A pesquisa descritiva é um dos meios formais adotado para realizar pesquisa científica, sendo que, a pesquisa descritiva é caracterizada pelo levantamento de dados e pela aplicação de entrevistas e/ou questionários (KNUPFER; MCLELLAN, 1996; KOTHARI, 2004). Além disso, a pesquisa descritiva pode ser considerada um passo prévio para encontrar fenômenos não explicados pelas teorias vigentes (LAKATOS; MARCONI, 2010; WASLAWICK, 2014). Desse modo, elaborar um questionário útil representa uma tarefa importante para a pesquisa descritiva.

Um dos meios de coletar dados é a aplicação de questionários. A aplicação de questionários permite a coleta de dados primários que são recolhidos pela primeira vez ou em replicações (aplicação do mesmo questionário, porém em datas ou para pessoas diferentes da aplicação anterior do questionário), e assim são de caráter original (KOTHARI, 2004). Questionários aplicados a um determinado problema de pesquisa podem ser de interesse do pesquisador, de modo que pode analisá-los antes de desenvolver ou aplicar seu próprio questionário.

Construir questionários não é uma tarefa fácil, portanto, aplicar algum tempo e esforço na sua construção pode ser um fator a ser investigado. O questionário deve ser muito bem organizado e conter uma ordem lógica para o entrevistado, evitando uma estrutura confusa e complexa, ou perguntas demasiadamente longas (LAKATOS; MARCONI, 2010; WASLAWICK, 2014). Questionário é uma ferramenta útil para diferentes comunidades, como Pesquisa em saúde (BOYNTON; GREENHALGH, 2004), Pesquisa educacional (ARTINO *et al.*, 2014) e Engenharia de software (MOLLÉRI; PETERSEN; MENDES, 2016). Pode ser uma fonte de obtenção de dados estatísticos para estudos de caso, fazendo comparações, sustentando argumentos e coletando opiniões. Algumas comunidades de pesquisa, como Hirsh Health Sciences [1], ADAI Library [2], RAND Health [3], IHSN [4], mantêm bancos de dados de questionário acessíveis. Eles incluem diferentes questionários que auxiliam profissionais e pesquisadores a analisar os resultados das questões ou formar uma estrutura para adicionar novas questões ou excluir questões inúteis. Eles também podem ser úteis para pesquisadores que buscam instrumentos de pesquisa validados e disponíveis. Nesse sentido, é interessante ter uma ferramenta que busca um questionário sem a necessidade de cadastro do usuário.

Considerando o ponto de vista do pesquisador, podem-se listar algumas van-

[1] https://researchguides.library.tufts.edu/c.php?g=249271&p=1659301
[2] http: //lib.adai. washington.edu/instrumentsearch.htm
[3] https://www.rand.org/health/surveys_tools.html
[4] http://www.ihsn.org/health-modules

tagens em se reutilizar questionários similares: (1) as perguntas de questionários já existentes poderiam ser incorporadas ao questionário de quem está produzindo um novo, com o intuito de ajudar o pesquisador a obter mais informações para sua pesquisa; (2) as respostas de perguntas semelhantes encontradas em outros questionários poderiam ser utilizadas por um pesquisador para fazer um comparativo entre a sua pesquisa e os trabalhos relacionados; (3) ao encontrar questionários similares, os mesmos podem corroborar com a pesquisa realizada pelo pesquisador, demostrando tendências ou mudanças ao longo do tempo em um determinado assunto de pesquisa.

Embora, até a presente data, não tenham sido encontrados trabalhos sobre recuperação ou ordenação de questionários de pesquisa, existem diversos trabalhos que vêm sendo aplicados na recuperação de informações não estruturadas no domínio de fóruns, conhecidos como *community question answering sites* (CQAs) (SRBA; BIELIKOVA, 2016). A semelhança entre uma página de CQA e um questionário de pesquisa está em se tratar de perguntas relacionadas a um determinado assunto.

Alguns exemplos são os trabalhos de CQAs cujo foco está na recuperação de perguntas, sem levar em consideração se fazem parte de um questionário (GRAPPY *et al.*, 2011; YANG; PIERGALLINI *et al.*, 2014; LIM; SACHAN; THING, 2013; HU; RUAN; SHAO, 2012). Outros fazem a análise de qualidade das respostas (WAMBSGANSS *et al.*, 2020; ABBASI; CHEN; SALEM, 2008; WEN; YANG; ROSE, 2014; ORTIGOSA; MARTÍN; CARRO, 2014; ROSENTHAL *et al.*, 2015; ABDUL-MAGEED; DIAB; KÜBLER, 2014; SMAILOVIC *et al.*, 2014; REFAEE; RIESER, 2014).

Já os trabalhos de *ranking* focam na ordenação por assunto (GUPTA; CARVALHO, 2019; ANWAR; ABULAISH, 2015; KIM; CAVEDON; BALDWIN, 2010; CHEN, R.-C. *et al.*, 2015). Também existem trabalhos que fazem a análise e mineração de dados (SCHMIDT; WEEDS; HIGGINS, 2020; DRINGUS; ELLIS, 2005; HUANG *et al.*, 2014; ABDELHAMID; AYESH; THABTAH, 2014; SADILEK *et al.*, 2016). E ainda pode-se citar trabalhos que coletam dados (*crawlers*) em sistemas CQAs (CHEN, D. *et al.*, 2017; MEUSEL; MIKA; BLANCO, 2014; AGICHTEIN *et al.*, 2008; KIM; HA, 2016). Vale ressaltar que, dada sua relevância, é crescente o número de trabalhos relacionados cujo foco é a recuperação e análise de informação em sistemas de CQAs (SRBA; BIELIKOVA, 2016).

De maneiral geral, tais trabalhos focam em perguntas do tipo aberta e não avaliam o questionário como um todo. Por outro lado, questionários de pesquisa contêm perguntas do tipo fechada (perguntas com opções de resposta), conforme pode ser observado na Figura 1. Contudo, não foram encontrados trabalhos que coletam, ordenam ou se preocupam em lidar com os questionários em si.

Um questionário de pesquisa pode ser considerado um documento de texto (perguntas) não estruturado. Pode-se então utilizar algum modelo de ordenação de documentos. Esses questionários de pesquisa possuem uma natureza mais específica,

contendo alternativas para os entrevistados responderem. Por exemplo, os questionários da Figura 1 contém perguntas com e sem alternativas, isso difere questionários de pesquisa de sistemas de CQA onde, usualmente, as perguntas não contém alternativas. Além disso, os trabalhos de recuperação de informação não consideram o agrupamento de perguntas. Portanto, a recuperação de questionários é ainda é um campo pouco explorado na área de recuperação de informação.

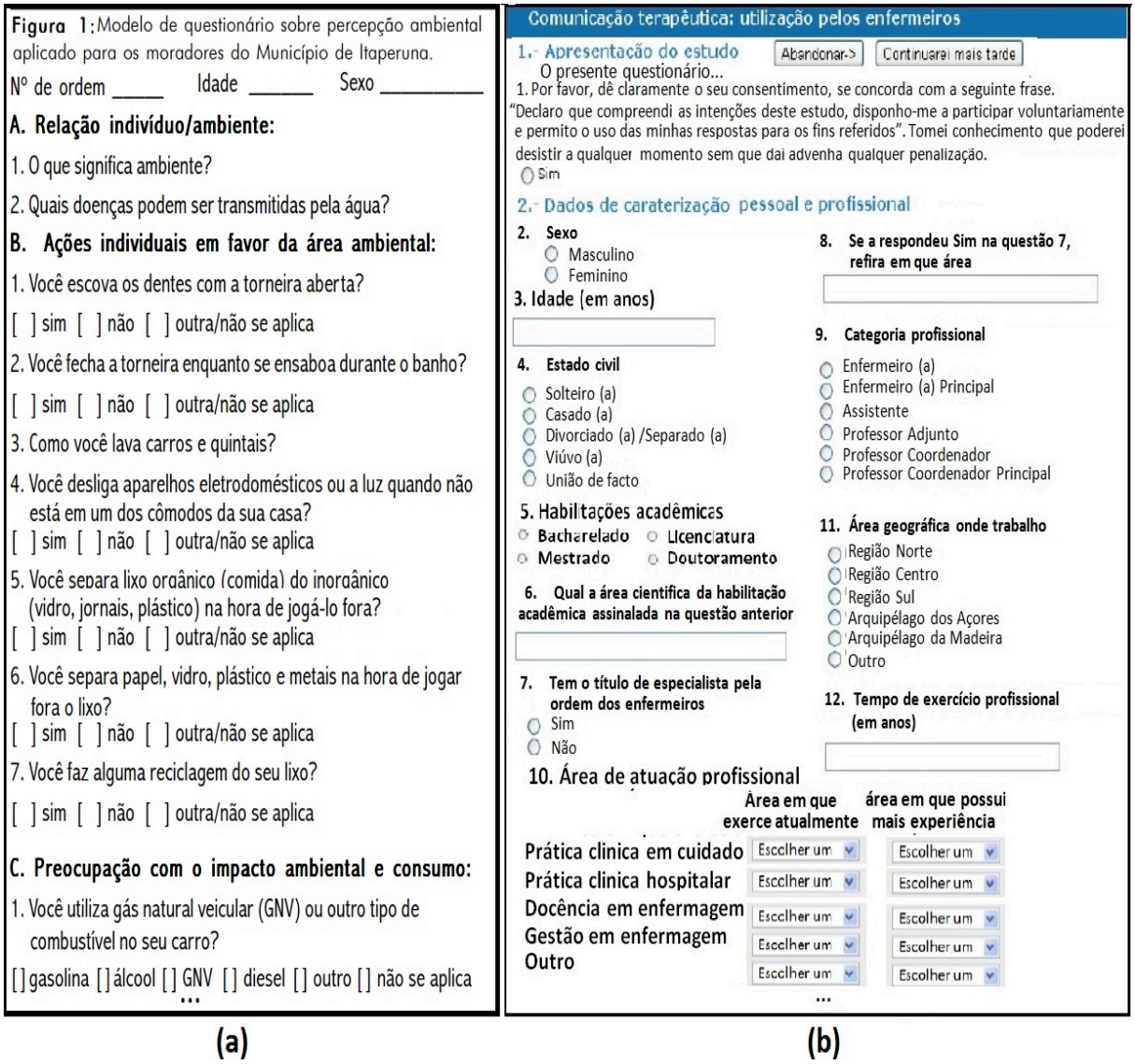

Figura 1: Modelo de questionário sobre percepção ambiental aplicado para os moradores do Município de Itaperuna.

Nº de ordem _____ Idade _____ Sexo _____

A. Relação indivíduo/ambiente:

1. O que significa ambiente?
2. Quais doenças podem ser transmitidas pela água?

B. Ações individuais em favor da área ambiental:

1. Você escova os dentes com a torneira aberta?
[ ] sim [ ] não [ ] outra/não se aplica
2. Você fecha a torneira enquanto se ensaboa durante o banho?
[ ] sim [ ] não [ ] outra/não se aplica
3. Como você lava carros e quintais?
4. Você desliga aparelhos eletrodomésticos ou a luz quando não está em um dos cômodos da sua casa?
[ ] sim [ ] não [ ] outra/não se aplica
5. Você separa lixo orgânico (comida) do inorgânico (vidro, jornais, plástico) na hora de jogá-lo fora?
[ ] sim [ ] não [ ] outra/não se aplica
6. Você separa papel, vidro, plástico e metais na hora de jogar fora o lixo?
[ ] sim [ ] não [ ] outra/não se aplica
7. Você faz alguma reciclagem do seu lixo?
[ ] sim [ ] não [ ] outra/não se aplica

C. Preocupação com o impacto ambiental e consumo:

1. Você utiliza gás natural veicular (GNV) ou outro tipo de combustível no seu carro?
[ ] gasolina [ ] álcool [ ] GNV [ ] diesel [ ] outro [ ] não se aplica

...

(a)

Comunicação terapêutica: utilização pelos enfermeiros

1.- Apresentação do estudo — Abandonar-> — Continuarei mais tarde

O presente questionário...

1. Por favor, dê claramente o seu consentimento, se concorda com a seguinte frase.

"Declaro que compreendi as intenções deste estudo, disponho-me a participar voluntariamente e permito o uso das minhas respostas para os fins referidos". Tomei conhecimento que poderei desistir a qualquer momento sem que daí advenha qualquer penalização.

○ Sim

2.- Dados de caraterização pessoal e profissional

2. Sexo
○ Masculino
○ Feminino

3. Idade (em anos)

4. Estado civil
○ Solteiro (a)
○ Casado (a)
○ Divorciado (a) /Separado (a)
○ Viúvo (a)
○ União de facto

5. Habilitações académicas
○ Bacharelado ○ Licenciatura
○ Mestrado ○ Doutoramento

6. Qual a área científica da habilitação académica assinalada na questão anterior

7. Tem o título de especialista pela ordem dos enfermeiros
○ Sim
○ Não

8. Se a respondeu Sim na questão 7, refira em que área

9. Categoria profissional
○ Enfermeiro (a)
○ Enfermeiro (a) Principal
○ Assistente
○ Professor Adjunto
○ Professor Coordenador
○ Professor Coordenador Principal

11. Área geográfica onde trabalho
○ Região Norte
○ Região Centro
○ Região Sul
○ Arquipélago dos Açores
○ Arquipélago da Madeira
○ Outro

12. Tempo de exercício profissional (em anos)

10. Área de atuação profissional

	Área em que exerce atualmente	área em que possui mais experiência
Prática clínica em cuidado	Escolher um	Escolher um
Prática clínica hospitalar	Escolher um	Escolher um
Docência em enfermagem	Escolher um	Escolher um
Gestão em enfermagem	Escolher um	Escolher um
Outro	Escolher um	Escolher um

...

(b)

Figura 1 – Exemplos com as 10 primeiras perguntas de questionários em pesquisas descritivas. (a) Questionário publicado no artigo de Villar *et al.* (2008), sobre a percepção ambiental entre os habitantes da região noroeste do estado do Rio de Janeiro. (b) Questionário publicado na tese Comunicação terapêutica em Enfermagem: utilização pelos enfermeiros (COELHO, 2015) e resultados publicados em artigo de Coelho (2015b).

1.1 MOTIVAÇÃO

Usualmente, quando um pesquisador faz uso de questionário em sua pesquisa, ele quer saber a opinião de outras pessoas sobre um determinado assunto. Porém, aplicar questionários pode envolver um custo elevado de tempo, dependendo de quantas pessoas o pesquisador quer obter a opinião. O tamanho do questionário também pode influenciar na vontade das pessoas o responderem, uma vez que um questionário muito grande exigirá um tempo considerável do entrevistado, o que pode acarretar em número pequeno de pessoas que estão dispostas a colaborar.

Dependendo do tipo de pesquisa a ser realizada, o pesquisador poderia averiguar se algum questionário semelhante já foi aplicado anteriormente. Caso encontre questionários semelhantes, poderia então utilizar apenas as perguntas que ainda não foram aplicadas. Por outro lado, existem situações nas quais não é interessante diminuir o questionário e sim aumentá-lo ou melhorá-lo. O pesquisador poderia analisar os questionários semelhantes, de forma a adicionar novas perguntas que passaram despercebidas na elaboração do questionário e que o pesquisador julgou relevante para sua pesquisa. Outra forma de melhoria pode ser na própria construção textual das perguntas, como alternativas não previstas em questões de múltipla escolha.

Outro aspecto interessante em se consultar questionários semelhantes já aplicados é poder analisar as respostas. Considere que um questionário recuperado seja apenas de perguntas abertas. A partir das respostas desse questionário pode-se agrupar as repostas repetidas ou similares. Assim, uma melhoria no questionário a ser aplicado seria a mudança da pergunta aberta para pergunta de múltipla escolha, facilitando, assim, que as pessoas respondam o questionário.

A Figura 1 mostra dois exemplos de questionários de pesquisa. No questionário do lado “a”, existem 19 perguntas utilizadas na pesquisa de Villar *et al.* (2008), sobre a percepção ambiental entre os habitantes da região noroeste do estado do Rio de Janeiro. Nota-se que o questionário possui duas questões abertas, isto é, duas questões onde o entrevistado é livre para responder como quiser. No entanto, as demais questões (17) são do tipo fechada, ou seja, o entrevistado deve escolher uma das opções previamente determinadas pelo pesquisador. Nota-se, portanto, uma predominância de perguntas que contêm alternativas. O mesmo pode ser observado no questionário do lado “b” da Figura 1, que mostra as 12 primeiras perguntas (de um total de 37), nas pesquisas realizadas por Coellho (2015a, 2015b) sobre Comunicação terapêutica em Enfermagem: utilização pelos enfermeiros. Neste questionário, pode-se observar que a pergunta número 10 é na verdade um conjunto de 12 perguntas do tipo fechada. Desse modo, pode-se afirmar que o lado “b” da Figura 1 contêm 19 perguntas com alternativas e apenas 4 do tipo aberta.

Vale ressaltar que pode ser possível realizar uma análise das respostas obtidas pelo pesquisador com as respostas dos questionários recuperados (LAKATOS; MAR-

CONI, 2010). Considerando-se o exemplo de um pesquisador de história social: esse pesquisador ao fazer um questionário sobre quem deve fazer as tarefas domésticas, pode comparar seus resultados com os resultados obtidos de questionários semelhantes de anos atrás, e então averiguar se haveria uma variação muito grande das respostas. Confirmaria a mudança de pensamento de década para outra? Mostraria tendências? (LAKATOS; MARCONI, 2010).

Encontrar questionários similares também pode ser útil para pesquisas que trabalham com dados secundários. Os dados secundários são aqueles que já foram coletados por outra pessoa e que já foram passados pelo processo estatístico. Desse modo, a pesquisa com dados secundários é um trabalho de natureza de compilação dos dados coletados (KOTHARI, 2004). Assim, quando o pesquisador optar por utilizar dados secundários, é importante averiguar quais dados já foram coletados por questionários semelhantes, a fim de corroborar com a pesquisa.

1.2 DEFINIÇÃO DO PROBLEMA

Considera-se que todo questionário é um conjunto Q = $\{(e, A)\}$, ou seja, um conjunto de perguntas, no qual (e, A) é uma pergunta. Para cada pergunta (e, A), e é o enunciado da pergunta e $A = \{a_1, a_2, ..., a_n\}$ é um conjunto de alternativas, tal que $e \in L$ e $a_i \in L$. L representa a linguagem escrita do idioma utilizado no questionário. L é passivel de tokenização, ou seja, pode ser dividida em palavras que pertencem à linguagem L_2; cada palavra em L_2 é chamada de *token*. A conversão em *tokens* é descrita na Seção 4.4.

Considerando a estrutura do questionário, o problema tem a seguinte instância:
Instância: um conjunto de questionários C, um questionário P.■

Dada a instância, tem-se a seguinte questão:
Questão: Qual é o valor da similaridade obtido da comparação de cada questionário Q $\in$ C com P.■

1.3 QUESTÃO DE PESQUISA

Conforme o problema descrito no presente trabalho, tem-se a seguinte questão de pesquisa:

Dado um conjunto de perguntas P, como analisar e ordenar por similaridade um ou mais conjuntos de perguntas que sejam semelhantes a P?

1.4 OBJETIVOS

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é o estabelecimento de uma abordagem de busca e ordenação de questionários por similaridade.

1.4.2 Objetivos Específicos

Para atingir o objetivo proposto, algumas metas foram definidas para serem executadas ao longo da elaboração da Tese. Assim, os seguintes objetivos específicos foram traçados:

1. Definir o modelo conceitual para recuperação de questionários.
2. Definir regras de similaridade, pois é necessário definir o quão similar os documentos são em relação à consulta.
3. Desenvolver um algoritmo para o cálculo da similaridade entre questionários.
4. Definir o formato de consulta, que seja adequado à busca por questionário, pois o uso direto de uma palavra-chave pode não trazer um conjunto de questionários que atendam à expectativa do usuário.
5. Desenvolver um algoritmo para realizar a ordenação dos questionários conforme a fórmula do cálculo de similaridade proposta.

1.5 CONTRIBUIÇÕES

A principal contribuição deste trabalho está no desenvolvimento de uma abordagem inédita de recuperação de questionários. Um protótipo é desenvolvido baseado na abordagem proposta para busca e ordenação de questionários (conjunto de perguntas) a partir de outro questionário, de modo que a ordenação é realizada com base no grau de similaridade entre questionários. Os resultados alcançados contribuem na área de recuperação da informação e de mineração de dados na Web, mais especificamente no campo de pesquisa de similaridade de textos.

A validação experimental do algoritmo foi realizada por meio de sua implementação e aplicação em uma base de dados real. Os resultados obtidos nos experimentos foram publicados em (SOUZA; DORNELES, 2017a,b, 2018, 2019a,b)[5], de modo a

[5] Analisando a Eficácia do Modelo Vetorial de Busca na Ordenação de Questionários(SOUZA; DORNELES, 2017a), publicado no SBSI 2017, QSMatching: an Approach to Calculate Similarity Between Questionnaires (SOUZA; DORNELES, 2017b) é um *short paper* publicado no iiWAS 2017, QSMatching vs Vector model: comparing effectiveness in questionnaires retrieval(SOUZA; DORNELES, 2018), publicado no SBSI 2018, Comparando a eficácia na recuperação de questionários: QSMatching vs Vector model vs Fuzzy (SOUZA; DORNELES, 2019a) é um artigo estendido do artigo

apresentar uma evidência da inovação científica deste trabalho na área de Recuperação de Informação.

De forma geral, pode-se especificar, pontualmente, as seguintes contribuições:

1. Organização de uma sub-área de pesquisa ainda não explorada, a de recuperação de questionário de pesquisa. Como pode ser observado através do levantamento do estado da arte, descrito no Capítulo 3, não foram encontrados trabalhos que realizam especificamente o cálculo de similaridade entre questionários;

2. Modelo de representação de questionários e consultas de forma a possibilitar o cálculo de similaridade entre questionários, independentemente se o parâmetro de consulta for um questionário, uma sentença ou uma palavra.

3. Especificação de um *crawler* focado para busca de questionários, a fim de construir uma base de dados para realização de experimentos que comprovem a eficácia da proposta apresentada na tese;

4. Definição de um *ground truth* para realização de testes de validação, com experimentos organizados de forma a envolver especialistas na área que comprovem a qualidade e corretude dos questionários coletados pelo *crawler*;

5. Organização de uma base de dados de questionários de forma a possibilitar experimentos na área, que envolvam esta e futuras pesquisas de recuperação de questionários.

1.6 ESTRUTURA DO TRABALHO

O trabalho está organizado em 6 capítulos, iniciando com esta introdução. O Capítulo 2 contém a fundamentação teórica, focando em conceitos necessários para o entendimento dos demais capítulos. O Capítulo 3 mostra os trabalhos relacionados juntamente com uma proposta de taxonomia baseada nos trabalhos pesquisados relacionados a sistemas CQAs. No Capítulo 4 é descrita a proposta da abordagem de recuperação de questionários. Uma proposta de um modelo em árvore para representar o questionário é descrita na seção 4.3 e a proposta de fórmula de similaridade entre questionários é definida na seção 4.5. Um resumo dos experimentos é descrito no Capítulo 5. Por fim, o Capítulo 6 descreve a conclusão.

publicado no SBSI2018 e publicado na revista iSys, Searching and ranking questionnaires: an approach to calculate similarity between questionnaires (SOUZA; DORNELES, 2019b), publicado no evento DocEng2019.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo mostra alguns conceitos básicos que norteiam esse trabalho. Inicialmente, é dada uma visão geral sobre questionários. Em seguida, são descritos os principais modelos de busca e, por fim, uma visão geral das técnicas utilizadas na área de recuperação de informação.

2.1 QUESTIONÁRIOS

Segundo Lakatos e Marconi (LAKATOS; MARCONI, 2010), questionário é um instrumento de coleta de dados, constituído por uma série ordenada de perguntas que devem ser respondidas sem a presença do entrevistador. Usualmente, o questionário deve ter um texto explicando a natureza da pesquisa, sua importância e a necessidade de obter respostas. O questionário deve ser feito de uma forma que desperte o interesse do entrevistado em responder as perguntas.

Um questionário é um meio para a obtenção de dados para a realização de pesquisa científica (SHEATSLEY *et al.*, 1983; HOVY *et al.*, 2000; ANDRADE *et al.*, 1999; LAKATOS; MARCONI, 2010; SOUZA MINAYO, 2011; VIEIRA, 2009; WASLAWICK, 2014). Ele é comumente utilizado na pesquisa descritiva como um esboço prático, sendo um passo prévio para encontrar fenômenos não explicados pelas teorias vigentes (LAKATOS; MARCONI, 2010; WASLAWICK, 2014). Uma opção, quando não é possível fazer pesquisa experimental, é utilizar questionários em pesquisas de levantamento, para ter a possibilidade de informações tabuladas de forma a serem tomadas conclusões sobre causas e efeitos (WASLAWICK, 2014).

Os questionários variam de acordo com as circunstâncias ou com o tipo de investigação. A elaboração de um questionário é uma tarefa que se aprende a realizar com a experiência (MALHOTRA, 2001). Na pesquisa descritiva, questionários são a fonte de obtenção de dados estatísticos que ajudam a realizar estudos de caso, comparativos, argumentações e coleta de opiniões. Um resultado da aplicação de questionários é a obtenção de evidências da validade dos resultados de uma pesquisa. Um exemplo de uso de questionários é na pesquisa quantitativa-descritiva dentro da pesquisa de campo, que se refere ao delineamento ou análise das características de fatos ou fenômenos, à avaliação de programas, ou ao isolamento de variáveis principais. Contudo, questionários utilizados em pesquisa de opinião, mesmo com a tabulação de dados e plotagem de gráficos, necessitam apresentar alguma informação nova para ter validade (WASLAWICK, 2014; LAKATOS; MARCONI, 2010; VIEIRA, 2009).

As perguntas de um questionário podem ser de diversos tipos (SHEATSLEY *et al.*, 1983; HOVY *et al.*, 2000; ANDRADE *et al.*, 1999; LAKATOS; MARCONI, 2010; SOUZA MINAYO, 2011; VIEIRA, 2009; WASLAWICK, 2014), todavia, o presente trabalho generaliza as classificações em apenas dois grupos de perguntas: discursiva

(aberta) e objetiva (fechada), ou seja, perguntas sem ou com alternativas, respectivamente.

As perguntas discursivas englobam as questões que são classificadas na literatura como do tipo aberto (SHEATSLEY *et al.*, 1983; LAKATOS; MARCONI, 2010). Perguntas abertas admitem uma quantidade maior de respostas diferentes entre os entrevistados, devido ao fato que cada entrevistado pode responder livremente. Uma pergunta aberta permite que se obtenha mais respostas não coincidentes entre os entrevistados, porque não há opções a serem selecionadas como resposta. As perguntas do item A do lado 'a' da Figura 1 e as perguntas 3, 6, 8 e 12 do lado 'b' da Figura 1 são exemplos de perguntas discursivas.

As perguntas objetivas são as que contêm alternativas e englobam dois tipos de classificação na literatura: a fechada e a de múltipla escolha. Nas perguntas fechadas, o pesquisador define as alternativas que podem ser a resposta do entrevistado, que deve optar por aquela que mais se ajusta às suas características, ideias ou sentimentos. Nas perguntas de múltipla escolha, são apresentadas várias alternativas e o pesquisado pode optar por mais de uma delas (respostas múltiplas), podendo ou não ter uma alternativa aberta que permita uma resposta diferente das alternativas apresentadas (SHEATSLEY *et al.*, 1983). As perguntas dos itens B, C e D do lado 'a' da Figura 1 e as perguntas 1, 2, 4, 5, 7, 9, 10 e 11 do lado 'b' da Figura 1 são exemplos de perguntas objetivas.

Cabe ressaltar que um questionário também pode ser utilizado em meios não científicos, podendo variar de importância, de pesquisas de mercado até como fazer uma festa de aniversário para um amigo em comum.

2.2 RECUPERAÇÃO DE INFORMAÇÃO

Conforme descrito em (MANNING; SCHÜTZE *et al.*, 1999), a eficácia da recuperação de informações tem impulsionado os motores de busca da web para novos níveis de qualidade onde a maioria das pessoas fica satisfeita na maior parte do tempo, e a pesquisa na web tornou-se uma fonte de informação padrão e a preferida para encontrar informação. O campo da recuperação de informações é um dos meios preferidos de acesso à informação pela maioria das pessoas (MANNING; SCHÜTZE *et al.*, 1999).

2.2.1 Pré-processamento

Uma etapa importante e que auxilia no processo de recuperação da informação é o pré-processamento, na qual um conjunto de ações transformadoras é aplicado, para garantir que a informação em formato texto se torne passível de ser analisada por algum algoritmo de recuperação de informação (GUIMARÃES; MEIRELES; ALMEIDA,

2019).

O texto livre tem infinitas possibilidades de escrita e pode existir erro de escrita como palavras duplicadas, falta de acentos, dentre outros. Desse modo, alguns trabalhos fazem a análise e correção do texto para deixá-lo em uma forma que possa ser trabalhado na recuperação de informação. Pode-se realizar uma análise em cada palavra do texto, assim como na análise léxica (AHMED *et al.*, 2015).

A verificação ortográfica também analisa se as palavras estão de acordo com algum idioma. Para ajudar na análise léxica ou na verificação ortográfica, pode-se remover letras repetitivas para retirar os erros de digitação. Em alguns trabalhos, a retirada de letras repetitivas é realizado na fase do pré-processemanto. Outra forma de ajudar essa análise computacional é a conversão dos caracteres para *lower-case* (BIYANI *et al.*, 2014; AHMED *et al.*, 2015; KIM; HA, 2016).

Os modelos clássicos de recuperação de informações pré-processam um documento de texto de forma a obter um conjunto de palavras-chave. O resultado do pré-processamento é um grupo de palavras (termos) selecionados, onde cada termo representa um conceito-chave ou tópico em um documento (BAEZA; RIBEIRO-NETO, 2011).

A *tokenização* é o processo de identificar *tokens* (ou palavras) nos documentos. A *tokenização* é a etapa de pré-processamento com a finalidade de extrair unidades mínimas de texto a partir de um texto livre (MANNING; SCHÜTZE *et al.*, 1999; FERREIRA, 2015).

O processo de *stemming* realiza a normalização linguística de acordo com um idioma. No caso do presente trabalho, o idioma é o português brasileiro. O algoritmo de *stemming* consiste basicamente da remoção de sufixos das palavras. Essa remoção é importante para que *tokens* semelhantes sejam agrupados em um mesmo radical (BAEZA; RIBEIRO-NETO, 2011; MANNING; SCHÜTZE *et al.*, 1999; FERREIRA, 2015).

A remoção dos *stopwords* é uma forma de limpeza do texto (SRIVIDHYA; ANITHA, 2010). Uma lista de *stopwords* é constituída pelas palavras de maior aparição em um documento texto e, normalmente, correspondem aos artigos, preposições, pontuação, conjunções e pronomes de uma línguagem. Devido a sua ocorrência alta no texto, sua presença na recuperação de informação apresenta um obstáculo (KANNAN; GURUSAMY, 2014).

O Quadro 1 mostra um exemplo de pré-processamento [1] utilizando como entrada a parte 'a' da Figura 1. Note no exemplo que a pergunta "**Você fecha a torneira enquanto se ensaboa durante o banho?**" após o pré-processamento é reduzida à "**voc fech torneir enquant ensabo banh**".

[1] O pré-processamento foi implementado pela ferramenta Lucene (APACHE. . . , 2020)

Quadro 1 – Pré-processamento

Documento	**Pré-processamento**
A. Relação indivíduo ambiente 1.O que significa ambiente? 2.Quais doenças podem ser transmitidas pela àgua? B. Ações individuais em favor da área ambiental: 1.Você escova os dentes com a torneira aberta? []sim []não []outra []não se aplica 2.Você fecha a torneira enquanto se ensaboa durante o banho?[]sim []não []outra []não se aplica 3.Como você lava os carros e quintais? 4.Você desliga aparelhos eletrodomésticos ou a luz quando não está em um cômodo da sua casa? []sim []não []outra []não se aplica 5.você separa o lixo orgânico (comida) do inogârnico (vidro, jornais, plástico) na hora de jogá-lo fora? []sim []não []outra []não se aplica 6.Você separa papel, vidro, plástico e metais na hora de jogar fora o lixo?[]sim[]não[]outra []não se aplica 7.Você faz alguma reciclagem de seu lixo? []sim []não []outra []não se aplica C.Preocupação com o impacto ambiental e consumo: 1.Você utiliza gás natural veicular (GNV) ou outro tipo de combustível no seu carro? []gasolina []álcool []GNV []diesel [] outro [] não se aplica	relaca, individu, ambient, 1, signif, ambient, 2, doenc, pod, ser, transmit, pel, agu, b, aco, individu, favor, are, ambiental, 1, voc, escov, dent, torneir, abert, sim, nao, nao, aplica 2, voc, fech, torneir, enquant, ensabo, banh, sim, nao, nao, aplica 3, voc, lav, carr, quint, 4, voc, deslig, aparelh, eletrodomest, luz, nao, est, comod, cas, sim, nao, nao, aplica 5, voc, sep, lix, organ, com, inogarn, vidr, jorn, plastic, hor, jog, lo, for, sim, nao, nao, aplica 6, voc, sep, papel, vidr, plastic, met, hor, jog, for, lix, sim, nao, nao, aplica 7,voc, faz, algum, reciclag, lix, sim, nao, nao, aplica c, preocup, impact, ambiental, consum, 1, voc, utiliz, gas, natural, veicul, gnv, tip, combust, no, carr, gasolin, alcool, gnv, diesel, nao, aplic

2.2.2 Modelos de Busca

O objetivo dos modelos de busca é permitir um mapeamento entre a consulta de um usuário e os itens no banco de dados de informações que atenderão a essa consulta (KOWALSKI; MAYBURY, 2006). A seguir é descrito brevemente os modelos clássicos (BAEZA; RIBEIRO-NETO, 2011).

2.2.2.1 Modelo *Booleano*

O modelo *booleano* é um modelo de recuperação simples baseado na teoria de conjuntos e álgebra *booleana*. Como resultado, o modelo é bastante intuitivo e tem semântica precisa. O modelo *booleano* considera que os termos da consulta estão presentes ou ausentes em um documento. Uma consulta *q* é composta de termos que podem ser ligados por três conectivos: *não*, *e*, *ou*. Portanto, uma consulta é essencialmente uma expressão *booleana* convencional entre a consulta e os documentos

(BAEZA; RIBEIRO-NETO, 2011).

O modelo *boleano* classifica o documento como relevante ou não relevante. Não há nenhuma noção de uma correspondência parcial entre a consulta e os documentos. Este critério de decisão binário sem qualquer noção de uma escala de classificação impede a boa qualidade da recuperação. Além disso, embora as expressões *booleanas* tenham semântica precisa, não é simples traduzir uma necessidade de informação em uma expressão *booleana*. A maioria dos usuários encontra dificuldade em expressar suas solicitações de consulta em termos de expressão *booleana*. Como consequência, as expressões *booleanas* formuladas pelos usuários são muitas vezes bastante simples (BAEZA; RIBEIRO-NETO, 2011).

Segundo Baeza-Yates e Ribeiro-Neto (2011) as principais vantagens do modelo booleano são o formalismo do modelo e a sua simplicidade. Por outro lado, as principais desvantagens são: (1) que não há ordenação, então o documento que o usuário realmente está procurando pode ser o último a ser visto; (2) e que formular consultas booleanas é complicado para a maioria dos usuários.

2.2.2.2 Modelo *Fuzzy*

A teoria dos conjuntos *fuzzy* lida com a representação de classes cujos limites não são bem definidos. A ideia principal é associar uma função de associação aos elementos da classe. Essa função retorna valores no intervalo [0, 1] com 0 correspondendo a nenhuma associaçãoo a uma determinada classe e 1 correspondente à participação plena. Assim, a participaçãoo em um conjunto *fuzzy* é uma noção intrinsecamente gradual. Como resultado, a correspondênncia de um documento com os parâmetros da consulta é aproximada (ou vaga). Isso pode ser modelado considerando que cada parâmetro de consulta define um conjunto *fuzzy* e que cada documento possui um grau de associação (geralmente menor que 1) nesse conjunto (BAEZA; RIBEIRO-NETO, 2011; KOWALSKI; MAYBURY, 2006).

O Lucene (APACHE. . . , 2020) implementa o modelo *fuzzy* baseado no algoritmo de distância de edição. Por exemplo, para fazer uma pesquisa *fuzzy* usando a palavra "dourado" como parâmetro de busca, os possíveis retornos são as palavras "ourado" e "dourar" que são semelhantes a ortografia "dourado".

2.2.2.3 Modelo Vetorial

O modelo vetorial é um dos modelos clássicos de recuperação de informação, onde cada documento é representado por um conjunto de *keywords* (termos indexados). Esses termos podem ser constituídos por uma palavra ou grupo de palavras consecutivas em um documento. Esse conjunto de termos é comumente obtido após o pré-processamento (ver Seção 2.2.1) dos documentos (*tokenização* e *stemming*) a fim de extrair tais termos (BAEZA; RIBEIRO-NETO, 2011).

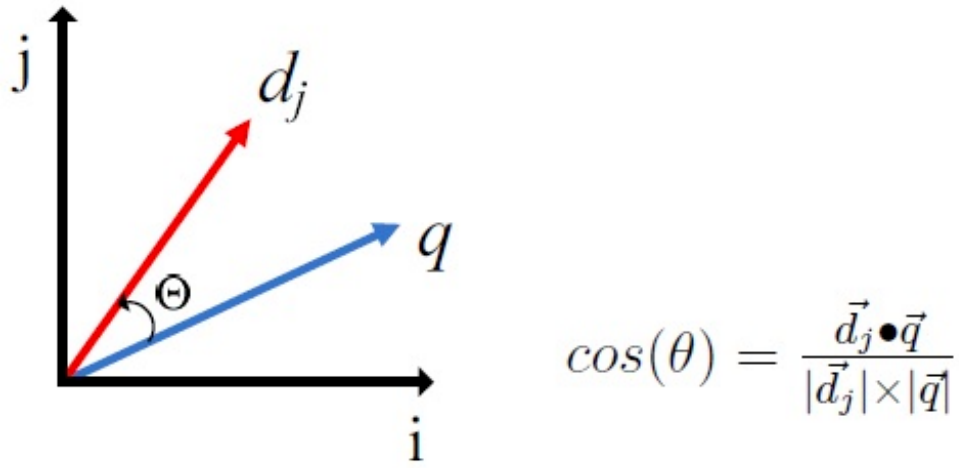

Figura 2 – Cálculo do Cosseno. Fonte:(BAEZA; RIBEIRO-NETO, 2011)

Um dos cálculos realizados na aplicação do modelo vetorial é o TF-IDF, usado para definição de pesos nos termos, onde TF é a frequência do termo no documento e IDF é o inverso da frequência do termo entre os documentos da coleção. Dessa forma, pode-se classificar os documentos por meio de atribuição de pesos para o índice de termos nas consultas. Os pesos dos termos são usados para calcular o grau de similaridade de cada documento com a consulta. Usualmente, é utilizado o cálculo do cosseno para definir o grau de similaridade entre a consulta e o documento. Assim, o cosseno é proporcional ao cosseno do ângulo entre o vetor que representa o documento e o vetor da consulta, conforme ilustrado na Figura 2 (BAEZA; RIBEIRO-NETO, 2011).

Cada termo da coleção é associado com uma unidade no vetor $\vec{k}_i$ em um espaço t-dimensional. Para cada termo $\vec{k}_i$ de um documento d_j é associado ao componente do vetor de termos $W_{i,j}$ *x* $\vec{k}_i$. Então o vetor do documento é dado por (MANNING; SCHÜTZE *et al.*, 1999; BAEZA; RIBEIRO-NETO, 2011):

$$\vec{d}_j = (w_{(1,j)}, w_{(2,j)}, ..., w_{(t,j)}) \tag{1}$$

onde t é o número de termos.

O tamanho do documento é dado pela norma do vetor (do próprio documento), calculado da seguinte forma (MANNING; SCHÜTZE *et al.*, 1999; BAEZA; RIBEIRO-NETO, 2011):

$$|\vec{d}_j| = \sqrt{\sum_{i}^{t} w_{i,j}^2} \tag{2}$$

O peso $w_{i,j}$ associado ao par (k_i, d_j) é não negativo e não binário. Os termos indexados são considerados todos independentes entre si. Eles são representados como vetores unitários de um espaço t-dimensional (t é o número total de termos indexados). A representação do documento d_j e uma consulta q são vetores ($\vec{}$) t-

dimensionais conforme Equação 1 e Equação 3 (MANNING; SCHÜTZE *et al.*, 1999; BAEZA; RIBEIRO-NETO, 2011):

$$\vec{q} = (w_{(1,q)}, w_{(2,q)}, ..., w_{(t,q)}) \tag{3}$$

Com os vetores do documento e da consulta é possivel realizar o cálculo do cosseno conforme a Figura 2.

2.3 MEDIDA DE SIMILARIDADE DE TEXTO

Na literatura existem várias maneiras de se medir a similaridade textual. Algumas medem textos longos tais como o TF-IDF, onde TF (*Term Frequency*) é a frequência de termos, sendo que quanto maior, mais relevante é o termo para descrever o documento, e IDF (*Inverse Document Frequency*) indica o termo que aparece em muitos documentos (BAEZA; RIBEIRO-NETO, 2011). Outra métrica é a *Co-Occurrence Frequency*, a qual representa a frequência de ocorrência, dentro de um texto, de dois termos seguidos em determinada ordem (ANWAR; ABULAISH, 2015). Ainda, com *Strings* longas, podemos citar *Normalized Google Distance* (NGD) (FAULKNER, 2014; CILIBRASI; VITANYI, 2007), *Wikipedia-based similarity metric* (WLM) (WITTEN; MILNE, 2008) e *Contingency Coefficient* (MANNING; SCHIITZE, 1999).

Por outro lado, também é possível medir textos (palavras) curtos, descrito na próxima Seção.

2.3.1 Medida de Similaridade de palavras

Uma pergunta pode ser considerada um pequeno conjunto de palavras; nesse caso, se considerarmos uma questão como uma unidade de comparação, pode ser interessante usar algoritmos de comparação de palavras. A seguir é descrito alguns algoritmos para calcular a similaridade entre palavras.

Uma medida utilizada na similaridade de palavras é a *edit distance* ou algoritmo de *Levenshtein*. Foi um dos primeiros algoritmos de comparação de *strings* e, até hoje, ainda é um dos mais usados (NAVARRO, 2001; GU; WANG; ZHAO, 2019). Ele considera o menor número de inserções, exclusões e substituições para transformar uma sequência na outra. Pontuações diferentes são atribuídas para cada operação possível: correspondência (casamento, igualdade de caracteres); desencontros; inserções, exclusões (NAVARRO, 2001; LEVENSHTEIN, 1966).

O algoritmo *Smith Waterman* é semelhante ao *Levenshtein*. O algoritmo realiza o alinhamento de *substring* encontrada entre duas *Strings* (palavras). Por exemplo, “questionário” e “questão” têm uma boa pontuação porque a *substring* alinhada é *quest* (SMITH; WATERMAN *et al.*, 1981).

O algoritmo *Jaro Metric* é baseado no número de caracteres iguais entre duas cadeias, e na semelhança da ordem em que essas duas cadeias são apresentadas (BILENKO *et al.*, 2003). Por exemplo, “questionário” e “questão” têm uma distância de 0,28.

O algoritmo de distância de *Hamming* tenta resolver o problema de pesquisa chamado “correspondência de *string* com *k* diferenças”. O algoritmo calcula a similaridade usando o número de bits que diferem entre duas sequências binárias ou o número de bits que precisam ser modificados para transformar uma sequência em outra (NAVARRO, 2001; CHAPMAN, 2007). Por exemplo, “questionário” e “questão” têm uma distância de 41 (bits).

O *Soundex Distance Metric* reduz cada *string* a um “Código *Soundex*”, que consiste em uma letra e três dígitos. Considera semelhantes todas as *strings* que têm o mesmo código. O objetivo desse método é transformar um nome em um código de 4 digitos para que sons semelhantes tenham esses 4 caracteres. O primeiro caractere é a primeira letra do nome e as próximas são substituídas por números (HALL; DOWLING, 1980). Por exemplo, “questionário” e “questão” são códigos diferentes.

A função de distância de *Covington* é usada para fazer comparações que levam em consideração se o termo comparado é uma vogal ou uma consoante. É uma espécie de comparação fonética superficial. Ele atribui pesos diferentes às substituições dos pares de segmentos e possui custos de inserção ou remoção independentes do contexto (KONDRAK, 2003). Por exemplo, “questionário” e “questão” têm distância de 12.

Um *n*-gram é o conjunto de todas as *substrings* que podem ser geradas a partir de uma determinada *string*. “*n*” (de *n*-gram) representa o tamanho dessas *substrings*. Exemplo (n-grama gerado para a sequência “questão” com $n = 3$): {# #q, #qu, ues, est, sta, tao, ao $, o $ $ }. Esse algoritmo pode ser usado como uma técnica de filtragem, cujo objetivo é descartar áreas onde não pode haver correspondência (FOSTER JR; EVANS, 2003). Por exemplo, “questionário” e “questão” têm 4 *substrings* iguais.

2.4 AVALIAÇÃO DE EFICÁCIA

Na recuperação de informação, a precisão e a revocação são algumas das principais medidas para avaliação da eficácia de sistema de recuperação de informação. A precisão é a relação entre o número de itens retornados e considerados relevantes com o número total de itens (KOWALSKI; MAYBURY, 2006), conforme a Equação 4:

$$Precisão = \frac{(número\ \ de\ \ itens\ \ relevantes\ \ recuperados)}{(número\ \ de\ \ itens\ \ recuperados)} \tag{4}$$

A revocação é a razão entre a quantidade de itens relevantes retornados e a quantidade total de itens relevantes presente na fonte de dados (KOWALSKI; MAY-

BURY, 2006), conforme a Equação 5:

$$Revocação = \frac{(número\ de\ itens\ relevantes\ recuperados)}{(número\ de\ itens\ relevantes)} \tag{5}$$

A partir de precisão e revocação calcula-se o *F-value* (ou F1), a qual é uma medida que usa uma média harmônica de precisão (P) e revocação (R) (KOWALSKI; MAYBURY, 2006; BAEZA; RIBEIRO-NETO, 2011), conforme a Equação 6:

$$F\text{-}value = \frac{(2 * P * R)}{(P + R)} \tag{6}$$

A média harmônica é sempre menor que a aritmética ou a média geométrica, e geralmente bem próxima do mínimo da precisão e da revocação (MANNING; SCHÜTZE *et al.*, 1999).

Outra métrica utilizada na comunidade de recuperação de informação é a *Mean Average Precision* (MAP), que fornece uma medida de qualidade entre os níveis de revocação (MANNING; SCHÜTZE *et al.*, 1999). MAP é a média do valor de precisão obtido para o conjunto de k documentos existentes após a recuperação de cada documento relevante. Ou seja, se o conjunto de documentos relevantes é dado por $q_j \in Q$ é $\{d_1, d_2, ..., d_{m_j}\}$ e R_k é o conjunto de resultados de recuperação classificados do resultado principal até recuperar o documento d_k, então o cálculo do MAP é dado por[2]:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} (\frac{1}{m_j}) \sum_{k=1}^{m_j} (Precisão(R_k)) \tag{7}$$

Quando um documento relevante não é recuperado, o valor da precisão na equação 7 é considerado 0 (MANNING; SCHÜTZE *et al.*, 1999).

A precisão e a revocação, amplamente utilizados, permitem apenas avaliações de relevância binária e podem ser fortemente influenciados por documentos relevantes encontrados no final do *ranking*. Como resultado, eles podem desfocar a distinção entre um modelo de recuperação de informação que recupera documentos altamente relevantes no topo do ranking e outro modelo que recupera apenas alguns documentos relevantes no topo do *ranking* (BAEZA; RIBEIRO-NETO, 2011). DCG (*Discounted Cumulated Gain*) atribui descontos para documentos não relevantes e pontuação para os documentos relevantes.

A partir do DCG é possível montar o gráfico NDCG@k (*Normalized Discounted Cumulative Gain at top k positions*), o qual é o desconto de Ganho Cumulativo Normalizado nas *k* primeiras posições. NDCG@k é uma métrica comumente usada para verificar a efetividade de um *ranking*, pois calcula o quão próximo um *ranking* está do seu *ground truth*. Formalmente é definido pela Equação 8, onde k é o número das

[2] $\sum$ símbolo de somatória. | | símbolo de módulo

primeiras posições. DCG@k refere-se ao ganho do *ranking* que se deseja avaliar e gDCG@k é o *ranking* esperado ou *ground truth*[3].

$$NDCG@k = \frac{DCG@k}{gDCG@k} \tag{8}$$

O MAE (*Mean absolute error*) é uma medida de erros entre as curvas de precisão e revocação. MAE pode ser calculado da seguinte maneira:

$$MAE = \frac{1}{R_p} \sum_{i=1}^{R_p} (p_i - e_i) \tag{9}$$

Onde, $(p_1, ..., p_i)$ são as precisões nos níveis de revocação $(1/R, ..., R_p/R)$, em que R_p é o número de documentos relevantes recuperados pelo sistema, e $(e_1, .., e_p)$ as precisões estimadas nos níveis de revocação dos dados (ASLAM; YILMAZ, 2005).

Quanto menor o valor do MAE, mais eficiente é o sistema de recuperação de informação.

[3] A abordagem para a avaliação do sistema de recuperação de informações gira em torno da noção de documentos relevantes e não relevantes. Essa decisão é conhecida como o *gold standard* ou *ground truth* (MANNING; SCHÜTZE *et al.*, 1999; BAEZA; RIBEIRO-NETO, 2011)

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os trabalhos relacionados a propostas de recuperação de questionários. Primeiramente, é apresentada a forma como o mapeamento sistemático de literatura (MSL) foi realizado. Posteriormente, são descritos os trabalhos que possuem relação mais próxima à proposta apresentada nesta tese.

3.1 MAPEAMENTO SISTEMÁTICO

Os trabalhos relacionados foram elencados por meio da execução de um mapeamento sistemático de literatura (MSL). Esse MSL tem como objetivo analisar e sintetizar a literatura existente sobre recuperação de questionários, e listar as técnicas existentes na busca e recuperação de questionários, tanto em bases de dados quanto na *web* (BIOLCHINI *et al.*, 2005). Foi realizado um MSL devido à necessidade de evidenciar os trabalhos existentes de recuperação de informação que consideram a estrutura de um questionário. As bases de dados utilizadas foram: ACM Digital Library [1], IEEEXplore [2], Science Direct [3] e Springer [4].

Quadro 2 – Descrição dos elementos *PICOC*[1] da Pesquisa

Critérios	**Descrição**
População	Quem aplica questionário
Intervenção	Recuperação de informação
Comparação	Técnicas da área de Recuperação de Informação
Resultado	Recupera questionários
Contexto	Na *web* ou base de dados disponíveis

Quadro 3 – Termos de busca

Critérios	**Termos**	**Sinônimos**
P1	*Similarity*	*Resemblance, comparison,kinship, propinquity*
P2	*Measurement*	*Calculation, computation,account, rate, calculus, tally, concretion, estimate, appraisal, reckoning, sum*
P3	*Questionnaire*	*FAQ – (Frequently Asked Questions), Quiz*
P4	*Question*	*Asking*

O MSL tem a seguinte pergunta de pesquisa: quais metodologias existem para calcular a similaridade entre questionários? A pergunta é derivada da definição dos elementos apresentados no Quadro 2 [5]. Para realizar a busca foram definidos os termos de busca nas bases de dados conforme o Quadro 3.

[1] `http://dl.acm.org`
[2] `http://ieeexplore.ieee.org`
[3] `http://www.sciencedirect.com`
[4] `http://www.springer.com`
[5] PICOC: Population (População), Intervention (Intervenção), Comparison (Comparação), Outcome

Quadro 4 – Exemplo de conulta da MSL

(similarity OR resemblance OR comparison OR kinship OR propinquity) AND (measurement OR calculation OR computation OR account OR concretion OR rate OR calculus OR estimate OR appraisal OR reckoning OR tally OR sum) AND (Questionnaire OR Quiz OR FAQ OR 'Frequently Asked Questions') AND (Question OR Asking)

Quadro 5 – Critério de Inclusão

Critério	**Descrição do Critério de Inclusão**
CI1	Trabalhos que tratam da recuperação de informação em documentos que contenham listas de perguntas (questionários)
CI2	O cálculo de similaridade deve ser entre questionários ou entre perguntas.

Quadro 6 – Critério de Exclusão

Critério	**Descrição do Critério de Exclusão**
CE1	Trabalhos que utilizam questionários para validar sua pesquisa.
CE2	Trabalhos que utilizam questionários para avaliar os sistemas de recuperação de informação.
CE3	Trabalhos que fazem *matching* da pergunta com as respostas.

Quadro 7 – Critério de Qualidade

Critério	**Descrição do Critério de Qualidade**
CQ1	Os questionários devem ser ordenados (*ranking*) conforme a similaridade.
CQ2	Os trabalhos devem avaliar a precisão e revocação.
CQ3	Os trabalhos devem comparar a eficácia entre os métodos.

O Quadro 4 mostra o exemplo de consulta em IEEEXplore (`http://ieeexplore.ieee.org`), cujo resultado foi de 68 trabalhos retornados em abril de 2020. Ao realizar a busca nas demais bases de dados, o total de trabalhos encontrados foi de 283. A próxima etapa é a leitura dos resumos dos 283 trabalhos com intuito de selecionar quais trabalhos atendiam os critérios de inclusão (Quadro 5) e exclusão (Quadro 6), e assim ter a lista de trabalhos a ser realizada a leitura completa do texto. O resultado foi de 53 trabalhos selecionados para a leitura. Além dos critérios de inclusão e exclusão, utilizou-se os critérios de qualidade do Quadro 7 para uma seleção mais apurada dos trabalhos selecionados para leitura completa.

O resultado do MSL é que não foram encontrados trabalhos que realizam especificamente o cálculo de similaridade entre questionários. Contudo, após a leitura, vinte trabalhos destacaram-se por estarem relacionados a similaridade entre perguntas ou entre respostas para uma determinada pergunta, os quais podem ser agrupados em *ranking*, *data mining*, qualidade e *crawling*.

(Resultado), Context (Contexto).

3.2 TRABALHOS EXISTENTES

Nesta seção, os trabalhos existentes são apresentados de acordo com uma classificação que associa as propostas ao foco principal da solução. Inicialmente, são apresentados trabalhos que focam em soluções de *ranking*, em seguida trabalhos que utilizam técnicas de *data mining* na elaboração de propostas, mais adiante são discutidas propostas que presam pela análise da qualidade das perguntas, e finalmente, trabalhos que realizam *crawling*, ou coletas, de questionários.

3.2.1 *Ranking*

Ordenar (*Rank*) as perguntas conforme o que usuário necessita é uma das linhas de pesquisa encontradas no MSL. Para realizar a ordenação das perguntas, deve-se calcular a similaridade das perguntas existentes com uma nova pergunta (comumente utilizada em diferentes fóruns). Também existe a possibilidade de ordenar as respostas de uma dada pergunta. Os trabalhos a seguir buscam verificar qual é a melhor técnica de ordenação de perguntas ou de respostas para uma determinada questão.

O trabalho de (GUPTA; CARVALHO, 2019) faz o *ranking* de perguntas de uma FAQ. É realizado o cálculo da similaridade da pergunta e suas respostas e selecionado o melhor resultado para a ordenação das perguntas, ou seja, se uma das respostas obtiver uma similaridade maior do que a sua pergunta, o restante é descartado e considerado apenas o melhor resultado. A similaridade é uma adaptação do cálculo do cosseno (ver Seção 2.2.2.3). Já a escolha entre qual similaridade a ser usada, da pergunta ou das respostas, é realizada via *machine learning*.

Learning-to-rank é a técnica utilizada por Chen *et al.* (2015) para usar semântica na ordenação das respostas de uma determinada pergunta. A proposta é que o usuário dê como entrada uma pergunta para o sistema, e então é realizada a classificação de possíveis respostas. Sendo assim, foi utilizado um método que utiliza conhecimento de senso comum na Internet para calcular parentesco semântico de textos arbitrários. A ideia-chave é a de representar textos como “uma mistura ponderada de um conjunto predeterminado de conceitos”. Esses conceitos são geralmente referidos como entradas de página da *Wikipedia*[6], e o parentesco semântico é computado utilizando-se medidas de similaridade do modelo vetorial. Para a aprendizagem foi utilizado o modelo *skip-gram*[7] com uma arquitetura de rede neural para calcular os vetores de palavras.

[6] https://www.wikipedia.org/

[7] Skip-gram tenta maximizar a classificação de uma palavra com base em outra palavra na mesma frase. Cada palavra atual como uma entrada para um classificador com camada de projeção e prever palavras dentro de um determinado intervalo antes e depois da palavra atual (MIKOLOV *et al.*, 2013; CHEN, R.-C. *et al.*, 2015).

Quadro 8 – Comparativo entre trabalhos de *ranking*

	Ground truth	**Métrica de Avaliação**	**Similaridade**
Gupta e carvalho(2019)	CQA SemEval	NDCG@k	Cosseno
Chen *et al*. (2015)	GOV2	NDCG@k	TF-IDF; Cosseno; Skip-Grama.
Anwar e Abulaish (2015)	Criou uma própria	MAP	Co-Occurrence Frequency; TF-IDF; Cosseno; PMI; Overlap; Dice; Jaccard; Chi-Square; LLR; Phi Coefficient; Contingency Coefficient.
Kim *et al*. (2010)	Map Task Corpus	DCG; *F-value*.	n-gram; TF-IDF.

O trabalho de Anwar e Abulaish (2015) testou um conjunto de técnicas para calcular a similaridade de texto com intuito de verificar se postagem nas redes sociais estão associados a grupos terroristas. A ideia é utilizar palavras de 'ódio' (como matar, exterminar, etc) como entrada para realizar a ordenação de possíveis grupos terroristas.

Históricos de *chats* são importantes fontes de informação em sistemas de CQAs. Kim *et al.* (2010) identificam e separam as perguntas das respostas do histórico de um diálogo, e então ordenam as respostas para uma nova pergunta, de acordo com o que foi extraído nos diálogos anteriores.

Observa-se que os trabalhos de *ranking* possuem três elementos necessários para obter uma solução: *ground truth*, métrica de avaliação e o cálculo de similaridade (Quadro 8). O *ground truth* ou padrão ouro (*gold standard*) é uma base que contém a informação de quais documentos são relevantes ou não a uma determinada consulta do usuário (MANNING; SCHÜTZE *et al.*, 1999). Nota-se que os trabalhos de Gupta e Carvalho (2019)[8], Kim *et al.* (2010) [9] e Chen *et al.* (2015) [10] utilizam bases de dados prontas. Já Answar e Abulaish (2015) criaram uma base de dados para usar como *ground truth*.

[8] CQA SemEval - http://alt.qcri.org/semeval2017/task3/
[9] Map Task Corpus - http://groups.inf.ed.ac.uk/maptask/maptasknxt.html
[10] GOV 2 - *http*://*ir*.*dcs*.*gla*.*ac*.*uk*/*test*$_c$*ollections*/*gov*2−*summary*.*htm*

Para avaliar a eficácia das abordagens de recuperação de informação são utilizadas métricas de avaliação (ver Seção 2.4). Os trabalhos de Gupta e Carvalho (2019) e Chen *et al.* (2015) utilizam a métrica NDCG@k para avaliar a eficácia das abordagens propostas. Answar e Abulaish (2015) utiliza a métrica MAP e o trabalho de Kim *et al.* (2010) utiliza as métricas DCG e *F-value*.

A similaridade pelo cálculo do cosseno é utilizada nos trabalhos de Gupta e Carvalho (2019), Chen *et al.* (2015) e Answar e Abulaish (2015). TF-IDF é utilizado nos trabalhos de Chen *et al.* (2015), Answar e Abulaish (2015) e Kim *et al.* (2010). O trabalho de Kim *et al.* (2010) utiliza também *n-gram*. Já o trabalho Chen *et al.* (2015) utiliza *skip-grama*. O trabalho de Answar e Abulaish (2015) compara vários tipos diferentes para calcular a similaridade (*Co-Occurrence Frequency; TF-IDF; PMI-point-wise mutual information; Cosseno; Overlap; Dice; Jaccard; Chi-Square; LLR-log likelihood ratio; Phi Coefficient; Contingency Coefficient*).

3.2.2 *Data mining*

Técnicas de *Data mining* podem ser utilizadas como abordagens para minerar uma determinada informação dentro de fóruns, como por exemplo, encontrar *links* que levem para páginas de *phishing* com intuito de ajudar os fóruns a retirar *links* que podem "prejudicar" seus usuários. Os trabalhos a seguir focam em minerar dados em sistemas que armazenam perguntas e respostas.

O trabalho de Schmidt, Weeds e Higgins (2020) utiliza a tarefa de *data mining* chamada de classificação[11] para mapear elementos de PICO (*Population, Intervention, Comparison, Outcome*) em um sistema de perguntas e respostas (SCHMIDT; WEEDS; HIGGINS, 2020).

Para avaliar o andamento de turmas de alunos, Dringus e Ellis (2005) propõem a utilização da tarefa de *data mining* chamada de descoberta de associação [12]. A estratégia tem como objetivo descobrir e construir representações alternativas para os dados subjacentes de fóruns assíncronos da discussão de alunos em uma disciplina. Desse modo, é possível interceptar a informação para um instrutor e assim extrair do fórum as informações visíveis e úteis. Então, são adicionados indicadores de participação temporais para mostrar como melhorar a capacidade do instrutor em avaliar o progresso de uma discussão em um fórum.

Na mesma linha de descoberta de regras de associação, Huang *et al.*(2014) verificam se os *posts* mais votados (*superposts*) contribuem em termos de qualidade e

[11] Segundo (GOLDSCHMIDT; PASSOS, 2005), na mineração de dados, a tarefa de classificação "consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram."

[12] Segundo (GOLDSCHMIDT; PASSOS, 2005), na mineração de dados, a tarefa de Descoberta de associação é aquela que "Abrange a busca por itens que frequentemente ocorram de forma simultânea em transações do banco de dados."

Quadro 9 – Comparativo entre trabalhos de *Data Mining*

	Métrica de Avaliação	**Técnica**
Schmidt, Weeds e Higgins (2020)	Precisão; Revocação	Redes neurais
Dringus e Ellis (2005)	Não há	Regras de Associação
Huang *et al*. (2014)	Não há	Regras de Associação
Yang *et al*. (2014)	MAP	Regras de Associação
Abdelhamid *et al*. (2014)	Precisão	Regras de Associação
Biyani *et al*. (2014)	Precisão; *F-value*	Árvores de decisão
Kim *et al*. (2016)	Não há	Regras de Associação

se têm algum impacto no restante dos alunos. A ideia é estudar os dados extraídos para verificar o comportamento e a contribuição dos alunos nos fóruns de aprendizagem colaborativa on-line.

Um problema que pode ocorrer em um fórum de discussão de uma disciplina é a sobrecarga de *posts*. Nesse sentido, Yang *et al*.(2014) utilizam a 'descoberta de associação' para analisar o espaço e recurso que efetivamente caracterizam o comportamento dos alunos no fórum, a fim de obter uma lista de tópicos de interesse para os alunos.

Ao entrar em uma lista de discussão on-line, o usuário procura obter respostas para seu problema específico. Porém, o usuário pode encontrar *links* para sites indesejáveis, tais como sites de *phishing*. Nesse sentido, Abdelhamid *et al*. (2014) usam a 'descoberta de associação' para atacar o problema de sites de *phishing* dentro de fóruns.

Para analisar subjetividade de tópicos do fórum on-line, Biyani *et al*. (2014) utilizam a tarefa de 'classificação'. A ideia é identificar temas subjetivos de discussão, como um tópico que procura a opinião das pessoas, pontos de vista, avaliações, especulações e outros estados particulares, e um tema não-subjetivo como um tópico que procura informações factuais.

Para a criação de bases de conhecimento, o trabalho de Kim e HA (2016) utiliza a tarefa de 'descoberta de associação' para analisar os comentários (*tweets*) dos usuários. Os comentários são associados a positivo ou negativo. Para obter informações sobre o usuário que escreveu os comentários obtidos, foi definida uma lista de palavras-chave que podem ser pistas para as identidades dos usuários, tais como idade, sexo e ocupação (KIM; HA, 2016). Finalmente, esses dados foram processados para formar informação semântica sobre as preferências de usuários em pequenas empresas e agregadas para fornecer informações úteis para os pedidos relevantes de proprietários ou sistemas de perguntas e respostas que prestam serviços de recomendação.

O Quadro 9 mostra as métricas utilizadas para avaliar os experimentos e as

técnicas de *data mining* utilizadas pelos trabalhos. O trabalho de Schmidt, Weeds e Higgins (2020) utiliza a técnica de redes neurais. Neste caso é uma rede neural de aprendizagem supervisionada, onde envolve a aprendizagem de uma função a partir de exemplos de suas entradas e saídas. No caso de ambientes completamente observáveis, um agente sempre poderá observar os efeitos de suas ações e, consequentemente, poderá usar métodos de aprendizagem supervisionada para aprender a prevê-los (RUSSELL; NORVIG, 2002; CARVALHO *et al.*, 2011; HAN; PEI; KAMBER, 2011; GOLDSCHMIDT; PASSOS, 2005).

Os trabalhos Dringus e Ellis (2005), Huang *et al*. (2014) Abdelhamid *et al*. (2014), Yang *et al*. (2014) e Kim *et al*. (2016) utilizam as 'regras de associação' como técnica de *data mining*. Regras de associação é uma técnica usada na construção de relações sob a forma de regras entre elementos. As regras de associação buscam relações entre os atributos dos elementos. O objetivo é definir regras fortes de acordo com alguma medida do grau de interesse (FERRARI; SILVA, 2017).

Biyani *et al*. (2014) utilizam a técnica de 'árvores de decisão'. Essa técnica de aprendizagem de máquina consiste em escolher atributos e intervalos de valores, ou outros critérios de decisão, da base de dados que comporão cada nível de nós da árvore e construir os ramos de forma que otimize algum critério de qualidade (FERRARI; SILVA, 2017).

Com relação as métricas utilizadas, a precisão foi utilizada pelos trabalhos de Schmidt, Weeds e Higgins (2020), ABDELHAMID *et al*. (2014) e Biyani *et al*. (2014). Revocação apenas no trabalho de Schmidt, Weeds e Higgins (2020). *F-value* no trabalho de Biyani *et al*. (2014). Os demais trabalhos não apresentaram métricas de avaliação.

3.2.3 Qualidade

A qualidade do conteúdo gerado por usuários nos fóruns varia drasticamente (AGICHTEIN *et al.*, 2008). Nesse sentido, existem trabalhos para identificar e separar os conteúdos que têm qualidade dos que não têm. Assim, o usuário que for pesquisar por algo que já foi respondido poderá encontrar as respostas de melhor qualidade sem perder tempo analisando respostas de baixa ou nenhuma qualidade. Os trabalhos enquadrados nesta categoria são aqueles que verificam a qualidade das perguntas ou das respostas às perguntas. Sendo assim, essa categoria foi subdividida em qualidade da pergunta e qualidade da resposta.

Os trabalhos que verificam a qualidade na resposta, ou seja, se os *posts* de uma linha de assunto estão coerentes com o tópico (pergunta) inicial, também verificam qual *post* tem a "melhor" resposta para o tópico (pergunta) proposto. Para classificar as respostas conforme sua qualidade, Dalip *et al.* (2013) propõem uma estratégia para avaliar as respostas de um fórum de maneira automatizada. A abordagem utiliza a função do cosseno, os vetores de entrada para o cálculo é obtido por pares (consulta e

Quadro 10 – Comparativo entre trabalhos de Qualidade em CQAs

	Métrica de Avaliação	**Similaridade**
Grappy *et al*. (2012)	Precisão; MAE.	Booleano; Vetorial.
Dalip *et al*. (2013)	NDCG@K.	Cosseno.
Aslay *et al*. (2013)	Precisão.	Soundex; NGD.
Molino *et al*. (2014)	Precisão; MAP.	Jaccard; n-gram; cosseno.
Wanbsganss *et al*. (2020)	Qualitativa.	-

resposta). A classificação de qualidade da resposta é obtida por *feedback* dos usuários.

Em outra linha de trabalho, a qualidade das respostas é inferida ao comparar o texto da resposta com o texto da própria pergunta. Nesse sentido, Grappy *et al.* (2011) separam as perguntas das respostas e em seguida utiliza a ferramenta *Lucene* (APACHE. . . , 2020) com a configuração para utilizar o modelo *boleano* em conjunto com o vetorial.

O trabalho de Molino e Aiello (2014) tenta inferir a qualidade das respostas por meio da análise de correspondência entre as respostas a uma pergunta. O trabalho calcula a similaridade entre as respostas e entre a pergunta e cada resposta. Desse modo, a resposta que tem similaridade alta com as respostas e com a pergunta é considerada de qualidade alta, já a resposta com similaridade baixa com as outras respostas e com a pergunta é considerada de qualidade baixa.

Verificar a qualidade da resposta considerando o *expertise* do respondente é realizado por Aslay *et al.* (ASLAY *et al.*, 2013). Assim, quando uma resposta é selecionada por quem perguntou como sendo a “melhor resposta” a uma questão particular, é feita a comparação com as outras respostas da mesma pergunta. O sistema sugere as respostas de melhor qualidade conforme o cálculo de similaridade das respostas em relação a resposta selecionada anteriormente como a melhor. As respostas não classificadas (ou postadas após a escolha da melhor resposta por quem perguntou) recebem um grau de qualidade de acordo com a similaridade em relação a que foi selecionada como a ‘melhor’ resposta.

O trabalho de Wanbsganss *et al.* (2020) desenvolveu um *chatbot* para responder pesquisas dos alunos, de forma a construir um diálogo, quase como uma entrevista qualitativa entre professor e aluno. Contudo, o trabalho não explica como foi desenvolvido o *chatbot* e faz apenas uma avaliação subjetiva dos resultados a partir dos comentários dos alunos.

O Quadro 10 mostra as métricas utilizadas para avaliar os experimentos e o tipo de cálculo para similaridade nos trabalhos de qualidade. O trabalho de Dalip *et al.*

(2013) utiliza o cálculo do cosseno para determinar a similaridade. Além do cosseno, Molino *et al.* (2014) utiliza em conjunto Jaccard e n-gram para determinar a similaridade. O trabalho de Aslay *et al.* (2013) utiliza o NGD (*Normalized Google Distance*) e soundex. O algoritmo que junta o *booleano* com o modelo vetorial (implementação do *Lucene*) é utilizado por Grappy *et al.* (2012). Já Wanbsganss *et al.* (2020) não utiliza algoritmos de similaridade, apenas compara as opniões dos usuários e não deixa claro a implementação do *chatbot*.

No que diz respeito a métrica de avaliação, a precisão é utilizada em Grappy *et al.* (2012), Aslay *et al.* (2013) e Molino *et al.* (2014). O MAE também é utilizado por Grappy *et al.* (2012). O MAP é utilizado por Molino *et al.* (2014). Dalip *et al.* (2013) utilizou o NDCG@k para avaliar seus resultados. Wanbsganss *et al.* (2020) optou por realizar um questionário avaliativo aos usuários que partiparam do experimento.

3.2.4 *Crawling*

Com o progresso da Internet, as informações estão cada vez mais acessíveis e espalhadas pelo mundo virtual. Para recuperar informação sobre algo específico pode-se recorrer ao uso de *crawlers*. Em sistemas de CQAs, o uso de um *crawler* é particularmente útil para coletar informações de diferentes fóruns e FAQs dentre outras mídias sociais, de forma a coletar e obter as respostas disponíveis para uma mesma pergunta. Os trabalhos descritos a seguir mostram as especificidades de encontrar e coletar conteúdo de sites CQAs, de forma a não ter dados duplicados ou irrelevantes.

O trabalho de Chen *at al.* (2017) busca responder perguntas cuja a resposta encontra-se na *Wikipedia*. O *crawler* busca as repostas tendo como base um conjunto de perguntas, de modo que é realizado o cálculo de similaridade entre cada pergunta com os artigos da *Wikipedia*. Então são selecionados os cinco melhores artigos para serem analisados os parágrafos, e na segunda rodada são extraídos os 5 parágrafos que melhor respondem a pergunta (CHEN, D. *et al.*, 2017).

As respostas a uma pergunta podem ser acessadas na web por meio de *links* em cascata. Sendo assim, o trabalho de Lim *et al.*. (2013) procura extrair *links* das postagens em fóruns. Para agrupar *links*, regras são usadas para identificar os *links* de uma determinada pergunta. O *crawler* agrupa os *links* por meio de palavras-chaves extraídas do *link* da pergunta. O algoritmo funciona em duas fases: a fase de treinamento e a fase de extração de conteúdo real. Durante a fase de treinamento, primeiro o algoritmo identifica palavra-chave comum para os *links* de referência sobre as páginas que contêm as perguntas. Então, o algoritmo encontra o caminho dominante para os *links* de referência e *links* entre as respostas para cada tipo de página e, posteriormente, identifica o melhor caminho para diferenciar entre as diferentes regiões de conteúdo na página. Por fim, o conteúdo é extraído em unidades, durante a fase de extração do conteúdo real.

Quadro 11 – Comparativo entre trabalhos de *crawling*

	Similaridade	**Métrica de Avaliação**
Chen *et al*. (2017)	TF-IDF; n-gram .	Precisão.
LIM *et al*. (2013)	-	Precisão. Revocação.
Hu *et al*. (2012)	-	Precisão. Revocação.
Liu *et al*. (2011)	Cosseno	-

A estrutura HTML de uma página pode ser usada para ajudar o *crawler* a extrair seu conteúdo. Desse modo, o *crawler* descrito em Hu *et al*. (2012) necessita saber a URL da página a ser coletada. Realiza a extração dos metadados e categoriza em cinco tipos de estrutura. A extração das mensagens são realizadas conforme o tipo da estrutura.

Liu *et al*. (2011) também usam a estrutura em HTML. O primeiro passo é a representação da página web, na qual é analisada a estrutura da árvore DOM da página e a informação visual dos nós da árvore é anexada. Depois se inicia a extração de registros das respostas a partir das subárvores detectadas na árvore DOM e, em seguida, todos os registros de resposta são extraídos através da remoção dos dados indesejáveis. Na sequência, na fase de extração direta, as páginas com as respostas de diferentes sites de fórum são inseridas, mas cada página deve conter vários registros de respostas. Já na fase de extração baseada em *wrapped*, as respostas são empacotadas (*wrapper*) pelo emprego direto da extração de uma página de exemplo. Em seguida, os registros de avaliação e as respostas são extraídos, com suas *wrappers* correspondentes.

O Quadro 11 mostra um comparativo dos trabalhos em relação ao uso de algoritmos de similaridade e métricas de avaliação. O trabalho de Chen *et al*. (2017) utiliza TF-IDF para selecionar os artigos da *Wikipedia* e n-gram para selecionar os parágrafos a serem extraídos. O cálculo do cosseno é utilizado no trabalho de Liu *et al*. (2011). Os trabalhos de Lim *et al*. (2013) e Hu *et al*. (2012) utilizam algoritmos que buscam caminhos (URL) dentro das páginas e não usam similaridade de texto.

A métrica da precisão como forma de avaliação é utilizada nos trabalhos de Chen *et al*. (2017), Lim *et al*. (2013) e Hu *et al*. (2012). A revocação também é utilizada nos trabalhos de Lim *et al*. (2013) e Hu *et al*. (2012).

3.2.5 Análise e Discussões

Após a análise dos trabalhos, observa-se que as métricas de avaliação comumente utilizadas são: NDCG@k, precisão e revocação. Outro ponto é a utilização do cálculo do cosseno ou do TF-IDF, portanto, testar o modelo vetorial para recuperação

de questionário é considerado como ponto de partida para realização de experimentos no presente trabalho.

Contudo, nos trabalhos relacionados, ainda não foi encontrado um trabalho que responda a questão: Dado um conjunto de perguntas, é possível analisar e ordenar por similaridade um ou mais conjuntos de perguntas?

Para responder a essa questão, o trabalho deve, primeiramente, atender ao formato de entrada, ou seja, deve reconhecer um questionário dentro de um documento de texto. O reconhecimento de questionários dentro de um documento não foi abordado pelos trabalhos relacionados. Então, como primeiro desafio tem-se: verificar se existe algum algoritmo, técnica ou abordagem que, ao ser aplicado, consiga reconhecer que a entrada é um questionário.

Deve-se levar em conta também que as perguntas dentro de um questionário podem variar de assunto e de objetivo, ou seja, em um questionário pode haver vários tópicos distintos. Por exemplo, um questionário de avaliação institucional de uma universidade pode variar de perguntas sobre o desempenho dos professores, da infraestrutura da instituição, dos serviços prestados e do conteúdo do curso. Assim, como segundo desafio, deve-se lidar com essas diferentes estruturas que um questionário pode ter na hora de fazer a busca por questionários similares. Será melhor tratar cada seção do questionário como um questionário diferente? Ou como um único questionário? Qual dos dois modos tem resultados melhores? Existe algum algoritmo, técnica ou abordagem que ao ser aplicado encontre os questionários similares?

Do mesmo modo, é necessário averiguar se os *crawlers* existentes coletam questionários no formato apropriado. Também é necessário definir um formato que seja mais adequado para o armazenamento dos questionários.

4 PROPOSTA DE SIMILARIDADE ENTRE QUESTIONÁRIOS

Este capítulo trata da definição da arquitetura desenvolvida para calcular a similaridade entre questionários. Uma visão geral da arquitetura proposta como solução para obtenção de um *ranking* de questionários por similaridade é ilustrada na Figura 3. A arquitetura tem como pressuposto que os questionários estão armazenados em um *dataset*. A finalidade da arquitetura é realizar a ordenação dos questionários por similaridade com o parâmetro de consulta fornecido por um usuário. O pré-processamento proposto é realizado nos questionários. De forma a facilitar o cálculo da similaridade, os questionários são convertidos em uma estrutura em árvore. Em seguida, é realizado o cálculo de similaridade dos questionários em relação a entrada do usuário e a consequente ordenação. Por fim, o ranking dos questionários é exibido ao usuário.

A seguir a proposta é detalhada, na Seção 4.1 é descrito um cenário exemplo. Em seguida, na Seção 4.2 é definido o modelo conceitual do questionário. Uma estrutura auxiliar para cálculo da similaridade é definida na Seção 4.3. O pré-processamento proposto para questionários é descrito na Seção 4.4. A arquitetura e as fórmulas para cálculo da similaridade são apresentadas na Seção 4.5. Por fim, um exemplo é apresentado na Seção 4.6, junto com os algoritmos.

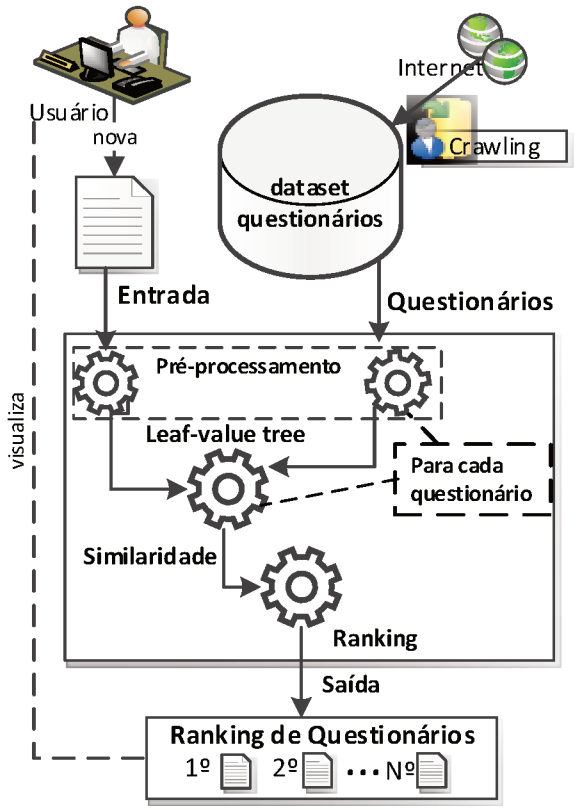

Figura 3 – Arquitetura proposta

Parâmetros da consulta	1- Qual foi o seu lucro ano passado? 2- Ano passado, você realizou alguma viagem de negócio?
Visualização parcial do Questionário da base de questionários	A qual faixa etária você pertence? a.()Abaixo de 17 d.()31 – 40 g.()Acima de 60 b.()17 – 20 e.()41 - 50 c.()21 – 30 f.()51 - 60 Qual é o seu nível de educação? a.( )Analfabeto / Primário incompleto f.( )Bacharelado b.( )Primário g.( )Mestrado c.( )Secundário incompleto h.( )Doutorado d.( )Secundário i.( ) Outro:_____ e.( )Curso técnico secundário Você vive com algum empresário ou pequeno empresário? a.()Sim b.()Não Qual foi a sua renda total em 2011? (sem deduções) Qual foi a renda familiar de sua casa em 2011? (sem deduções) Qual é a sua nacionalidade? Em que tipo de moradia você vive? a.()Morando com parentes e.()Sem teto b.()Morando com um parente idoso f.()Arrendada c.()Atualmente no sistema carcerário g.()Alugada d.()Casa ou apartamento de propriedade da família Qual é o seu estado civil? a.()Casado(a) c.()Viúvo(a) b.()Solteiro(a) d.()Outra: ___ Quando você viaja, qual é o motivo? a.()Turismo. b.()Negócio.

Figura 4 – Exemplo de consulta e questionário similares

4.1 CENÁRIO EXEMPLO

Na área de pesquisa qualitativa, são comumente usados questionários para realizar a análise de algum trabalho (SOUZA MINAYO, 2011). Um pesquisador que tenha elaborado um questionário para que as pessoas analisem seu trabalho pode então verificar se existe outro questionário de pesquisa semelhante. Desse modo, o pesquisador pode incrementar o seu próprio questionário com perguntas de questionários anteriores que são semelhantes ao seu. Outra possibilidade é a comparação das respostas do questionário elaborado pelo pesquisador com as respostas de questionários recuperados. Assim, pode-se verificar as tendências e diferenças entre trabalhos similares.

Esta seção descreve um cenário de exemplo para o problema tratado neste trabalho. A Figura 4 apresenta um conjunto de perguntas que pode ser usado como consulta a um repositório de questionários. Supõe-se que um pesquisador deseja buscar questionários que possuam perguntas tais como os "parâmetros de busca" apresentados na Figura 4. A ideia é verificar se existem questionários similares na área temática de finanças, especificamente em finanças pessoais[1]. Nesse contexto, é necessário então realizar uma busca por similaridade a repositórios de questionários já existentes[2]. A partir daqui, por questões de facilidade, "parâmetros da consulta" são chamados apenas de "consulta" e "questionário de pesquisa da base de questionários" é chamado apenas de questionário.

Considerando, então, o exemplo apresentado na Figura 4, uma primeira análise a ser feita é em relação ao tamanho do questionário. Levando em consideração apenas o número de perguntas, verifica-se que o questionário é constituído por nove perguntas enquanto a consulta é constituída por duas. Neste caso, seria possível supor que o questionário está 7 perguntas distante em relação a consulta. Porém, utilizar apenas o número de perguntas para obter o quão um questionário é diferente (distante) do outro e, por consequência, inferir a similaridade entre eles, é uma estratégia ingênua, uma vez que dois questionários com o mesmo número de perguntas podem ser completamente diferentes. Descartando a estratégia da contagem simples da diferença do número de perguntas entre o questionário e a consulta, propõe-se fazer uma análise pergunta por pergunta, de forma a procurar quais perguntas da consulta são similares em comparação com as perguntas do questionário. Desta forma, considera-se hipoteticamente uma função *compara(c, q)*, onde *c* é a consulta e *q* o questionário, as seguintes comparações seriam realizadas:

Comparação 1 *compara("Qual foi o seu lucro ano passado?", q)*: ao realizar uma comparação da pergunta 1 da consulta com as perguntas do questionário, o pesquisador encontra duas perguntas parecidas no questionário. Uma pessoa, nesse caso, o pesquisador, pode inferir que as perguntas "Qual foi a sua renda total em 2011? (sem deduções)" e "Qual foi a renda familiar de sua casa em 2011? (sem deduções)" são similares à pergunta 1. Note que, um dos fatores das perguntas serem ditas similares é que as palavras lucro e renda nas perguntas indicam que o entrevistado deve informar uma quantia em dinheiro. Outro fator é a questão da temporalidade: as três perguntas são referentes ao ganho de dinheiro em um ano. Desse modo, conseguir computacionalmente obter tal inferência, (ou seja, que as perguntas são similares) é um dos problemas a serem resolvidos. Observe que a pergunta "Qual foi a sua renda total em 2011? (sem deduções)" é mais similar à pergunta 1 da consulta do que a

[1] Um exemplo de uso de questionários de pesquisa na área temática de finanças pode ser visto em Claudino *et al*. (CLAUDINO; NUNES; SILVA, 2009)

[2] Considera-se, neste trabalho, quaisquer repositórios de questionários, sejam privados, públicos, disponíveis na Web, etc.;

pergunta “Qual foi a renda familiar de sua casa em 2011? (sem deduções)”. Embora uma pessoa consiga fazer tal afirmação, ainda é difícil identificar o quão similares são as perguntas, evidenciando outro problema a ser resolvido, que é medir a similaridade entre perguntas;

Comparação 2 *compara("Ano passado, você realizou alguma viagem de negócio?", q)*: ao realizar uma comparação da pergunta 2 da consulta com as perguntas do questionário, o pesquisador encontra uma pergunta parecida no questionário. No questionário, existe a pergunta “Quando você viaja, qual é o motivo?” com as alternativas “Turismo” e “Negócio”, de modo que, uma possível informação obtida é que o entrevistado fez uma viagem de negócio. Vale ressaltar que, tal conclusão foi possível ao ser considerada a pergunta em conjunto com suas alternativas. Observe que, o grau de similaridade pode ser obtido realizando uma análise nas suas respectivas opções de resposta.

Neste cenário, após a comparação das 2 perguntas da consulta com as 9 perguntas do questionário, obteve-se como resultado: 3 perguntas do questionário são similares às duas perguntas da consulta, portanto, o questionário pode ser dito como relevante. Vale ressaltar que, embora as perguntas sejam formuladas de maneiras diferentes, a consulta e o questionário verificam o ganho monetário pessoal e se o entrevistado fez viagem de negócio. Contudo, ainda fica a questão: o quão similar o questionário é em relação à consulta?

Um problema constatado é que palavras-chaves para o entendimento da pergunta podem estar nas alternativas, e portanto, afetar o cálculo de similaridade entre a consulta e o questionário. Por exemplo, a pergunta *“Você vive na casa dos seus pais?”* como sendo um parâmetro de consulta e que o questionário contenha a seguinte pergunta *“Em que tipo de moradia você vive?”* com as seguintes alternativas *“(a)Alugada”, “(b)Própria”, “(c)Sem Teto”, “(d)Com os pais”*. É possível notar que, a pergunta do questionário em conjunto com a alternativa *“Com os pais”* é fortemente similar a pergunta da consulta. Por outro lado, ao analisar apenas as perguntas sem as alternativas, a mesma conclusão de que são fortemente similares não é possível. Dessa forma, realizar uma análise de cada par (pergunta, alternativa) em relação à consulta pode indicar uma eficácia maior na recuperação de questionários.

4.2 MODELO

Nesta seção é descrita a proposta de um modelo conceitual de representação de questionários de pesquisa, apresentado na Figura 5. O modelo representa um questionário que é composto por um conjunto de perguntas. Cada pergunta é associada a um tipo (aberta ou fechada). A pergunta poderá conter alternativas apenas se o tipo da pergunta for fechada. Cada alternativa representa uma opção de resposta para o entrevistado.

Nesse modelo, a raiz representa o título do questionário, o segundo nível é composto pelos nós que representam as perguntas, e, no último nível, as folhas representam as alternativas (caso existam). No exemplo da parte B da Figura 5 tem-se que o título do questionário é "percepção ambiental", o qual se encontra na raiz da árvore. A pergunta "qual o tipo de combustível que você utiliza no seu carro?" encontra-se no nó do segundo nível da árvore. Por fim, as alternativas "a.gasolina; b.álcool; c.GNV; d.diesel; e. não se aplica" encontram-se nas folhas da árvore.

O diagrama de classes da Figura 6 ilustra, de forma resumida, uma possível implementação do modelo proposto. Assim, considera-se que o questionário tem no mínimo uma pergunta e que a pergunta pode ou não ter alternativas.

Embora o questionário seja representado como uma árvore de até três níveis, na maioria dos casos, a entrada do usuário pode ser uma simples palavra-chave. Por exemplo, a Figura 7 apresenta quatro possibilidades diferentes para formular uma consulta: (i) um questionário (consulta "A"); (ii) uma pergunta fechada (consulta "B"); (iii) uma pergunta aberta (consulta "C"); ou (iv) ou apenas uma palavra-chave (consulta "D"). Ao representar as consultas como árvores, como mostrado na coluna "árvore" da Figura 7, as consultas "C" e "D" possuem apenas um nível (somente a raiz), a consulta "B" possui dois níveis (primeiro nível com a questão e o segundo nível com as alternativas), e a pergunta "A" tem a mesma estrutura do questionário da Figura 5. Então, antes de verificar a similaridade entre um questionário com a consulta, propõe-se convertê-

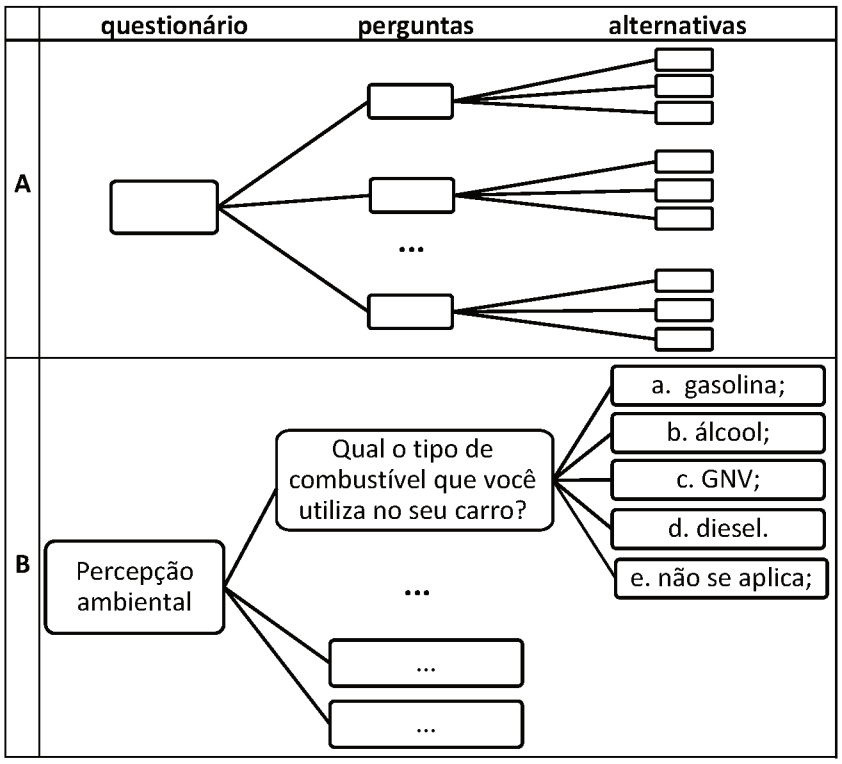

Figura 5 – Parte A: Modelo de dados; Parte B: exemplo do modelo

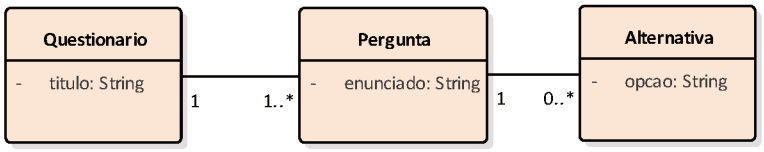

Figura 6 – Diagrama de classes para questionários

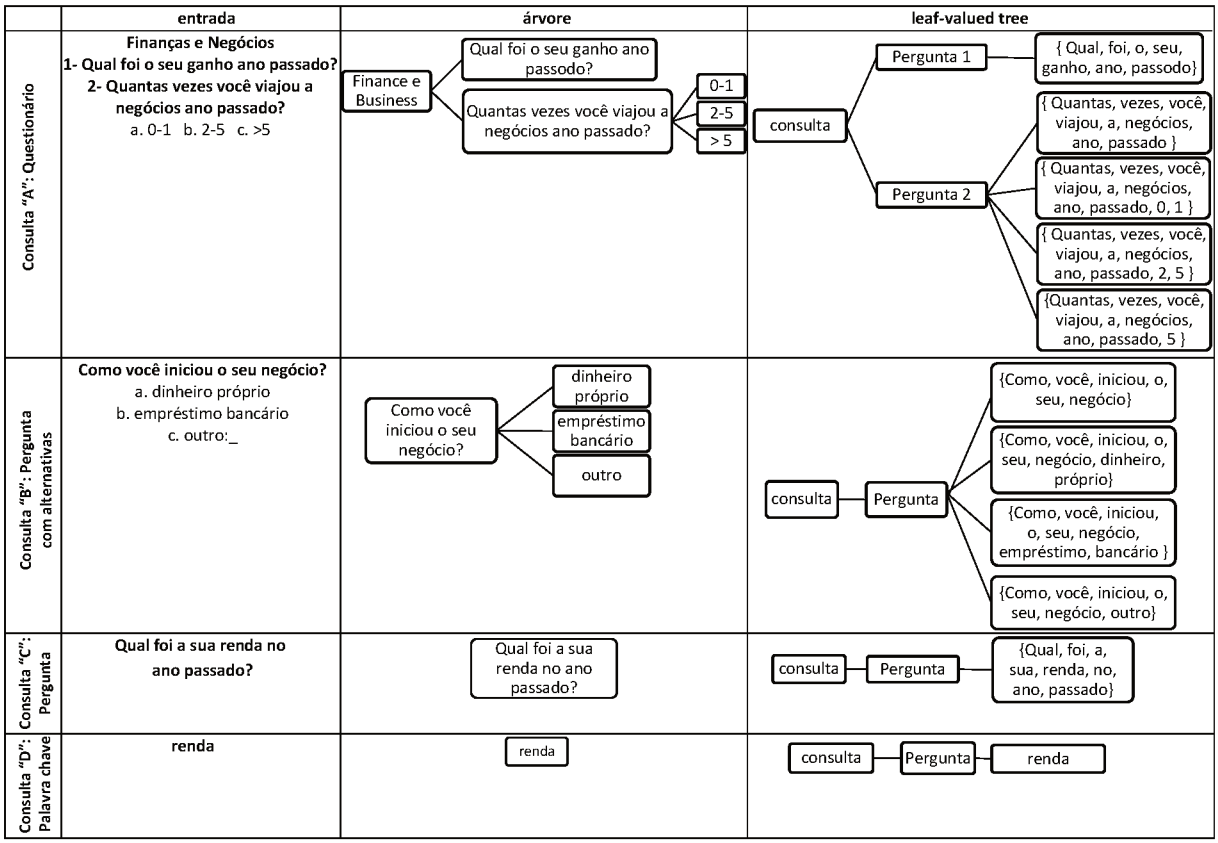

Figura 7 – Exemplos de queries

los, tanto a entrada quanto os questionários do *dataset*, em uma estrutura em árvore (*leaf-valued tree*) que tenha o mesmo número de níveis. Desse modo, a informação a ser usada no cálculo de similaridade está no mesmo nível (nas folhas). Essa conversão pretende facilitar o cálculo de similaridade entre a consulta e o questionário.

4.3 LEAF-VALUED TREE

De acordo com (PICARD, 1980), um questionário pode ser representado por um grafo conectado. Partindo desse princípio, propõe-se converter um questionário em uma árvore chamada de *leaf-valued tree*. Uma *leaf-valued tree* é uma representação interna de uma consulta ou um questionário e é projetada para ajudar no cálculo da similaridade. Em uma *leaf-valued tree*, a raiz (primeiro nível) e os nós (segundo nível) são pontos de referência do caminho que o algoritmo de cálculo de similaridade

deve seguir para chegar às folhas. No último nível, cada folha contém um conjunto de palavras. O cálculo da similaridade é realizado com o conteúdo das folhas.

A Figura 7, consulta “A”, mostra um exemplo de um questionário “Finanças e Negócios”, onde na última coluna é o questionário convertido em uma *leaf-valued tree*. É possível notar que a Pergunta 1 gera 1 folha e a Pergunta 2 gera 4 folhas. A primeira folha referente ao nó da Pergunta 2 contêm as “palavras” da pergunta sem suas possíveis respostas e as outras três folhas contêm a combinação das “palavras” da pergunta com as “palavras” de cada resposta possível.

As *leaf-valued trees* de um questionário ou de uma consulta são estruturadas da mesma maneira, mesmo que seja uma simples palavra-chave. Neste sentido, quando uma consulta não é um questionário inteiro (Consultas “B”, “C” e “D” na Figura 7), a *leaf-valued tree* também é estruturada como uma árvore de 3 níveis. No exemplo da consulta “B” é necessário adicionar apenas uma “raiz” para formar a *leaf-valued tree*. Nos exemplos das consultas “C” e “D” é necessário adicionar o nó correspondente à pergunta e a raiz.

Desse modo, tanto a estrutura da consulta quanto dos questionários estarão no mesmo formato na hora de calcular a similaridade. A conversão para *leaf-valued tree* é realizada de forma simultânea com o pré-processamento do texto. A seguir é explicado o algoritmo de conversão de um questionário em uma *leaf-valued tree*.

4.3.1 Algoritmo de conversão

Conforme o modelo da Figura 5, e de acordo com a definição de questionário por Picard (1980), um questionário pode ser representado por um grafo. Tal grafo pode ser simplificado de modo a conter apenas os enunciados das perguntas e das alternativas. Essa simplificação facilita a conversão de um questionário em uma *leaf-valued tree*.

O Algoritmo 1 é utilizado para converter o questionário em uma *leaf-valued tree*. Considera-se que todo questionário é um conjunto $Q = \{(e, A)\}$, ou seja, um conjunto de perguntas, no qual (e, A) é uma pergunta. Para cada pergunta (e, A), e é o enunciado da pergunta e $A = \{a_1, a_2, ..., a_n\}$ é um conjunto de alternativas, tal que $e \in L$ e $a_i \in L$. L representa a linguagem escrita do idioma utilizado no questionário. L é passível de tokenização, ou seja, pode ser dividida em palavras que pertencem à linguagem L_2; cada palavra em L_2 é chamada de *token*.

O Algoritmo tem como entrada o questionário que será convertido em uma *leaf-valued tree*. Na linha 2, a estrutura que conterá a *leaf-valued tree*, chamada de Lvt, é inicializada. O algoritmo então realiza a tokenização de cada enunciado (linha 4) das perguntas do questionário (T é a função de tokenização). Na linha 5, inicia-se a simplificação de uma questão. Para cada alternativa da pergunta, é realizada a tokenização da alternativa e a união com os *tokens* do enunciado da pergunta, e adiciona-se o subconjunto do enunciado mais a alternativa em q’ (linha 7). A Lvt

Algoritmo 1 Gerar *leaf-valued tree*

```
1: Entrada: Q
2: Lvt ← {}
3: para cada (e, A) ∈ Q faça
4:     t'← T(e)
5:     q' ← {t'}
6:     para cada a ∈ A faça
7:         q' ← q' ∪ {t' ∪ T(a)}
8:     fim para
9:     Lvt ← Lvt ∪{{q'}}
10: fim para
11: return Lvt
```

conterá cada conjunto de *tokens* do enunciado mais as alternativas (linha 9). Por fim, o algoritmo retorna o questionário na forma de uma *Leaf valued-tree* (Lvt).

4.3.1.1 Análise de complexidade

O Algoritmo 1 mostra o processo de conversão um questionário em uma *leaf-valued tree*. A conversão é executada no tempo $O(|e|.|Q|.|A|)$, já que a tokenização na linha 4 (T(e)) demora tempo $O(|e|)$, considerando uma tokenização através de um autômoto finito determinístico (*deterministic Finite automaton*)), ($|e|$ é a quantidade de simbolos do maior enunciado aceito; para cada uma das passagens pelas alternativas do conjunto A (linhas 6-8) demanda o tempo $O(|A|)$, e para cada passagem pelas perguntas nas linhas 3-10 demora o tempo $|e|+|e|.|A| = O(|e|.|A|)$, como são $|Q|$ passagens, o custo total é $O(|e|.|Q|.|A|)$.

4.4 PRÉ-PROCESSAMENTO

A proposta para o pré-processamento de texto em questionários é a remoção de determinados termos (palavras) usados em perguntas e em alternativas. Ao usar a árvore de 3 níveis, já estão identificadas quais sentenças representam perguntas e quais sentenças representam alternativas. Na *leaf-valued tree*, o texto do enunciado da pergunta é sempre a primeira folha de cada nó. O objetivo é melhorar a eficácia da busca e ordenação dos questionários sem alterar o cálculo de similaridade. Figura 8 é um exemplo da *leaf-valued tree* com e sem o pré-processamento. As palavras filtradas nas perguntas são as seguintes:

Pronomes interrogativos, tal como “onde” são comumente usados nas perguntas. Na amostra de 510 (9996 perguntas) questionários que foram coletados da Web, há pelo menos um pronome interrogativo (6394 perguntas contêm pronomes interrogativos, totalizando 63,96% das perguntas).

Verbos auxiliares, tais como “ser” e “estar” também são comumente usados nas perguntas, especialmente nas perguntas cujas as respostas são usualmente “SIM”

Quadro 12 – Pré-processamento comumente usado vs pré-processamento proposto

Entrada	Resultado do Pré-processamento	
	Clássico	Proposto
Onde você investe?	{ond, voc, invest}	{invest}
Onde você almoça?	{ond, voc, almoc}	{almoc}

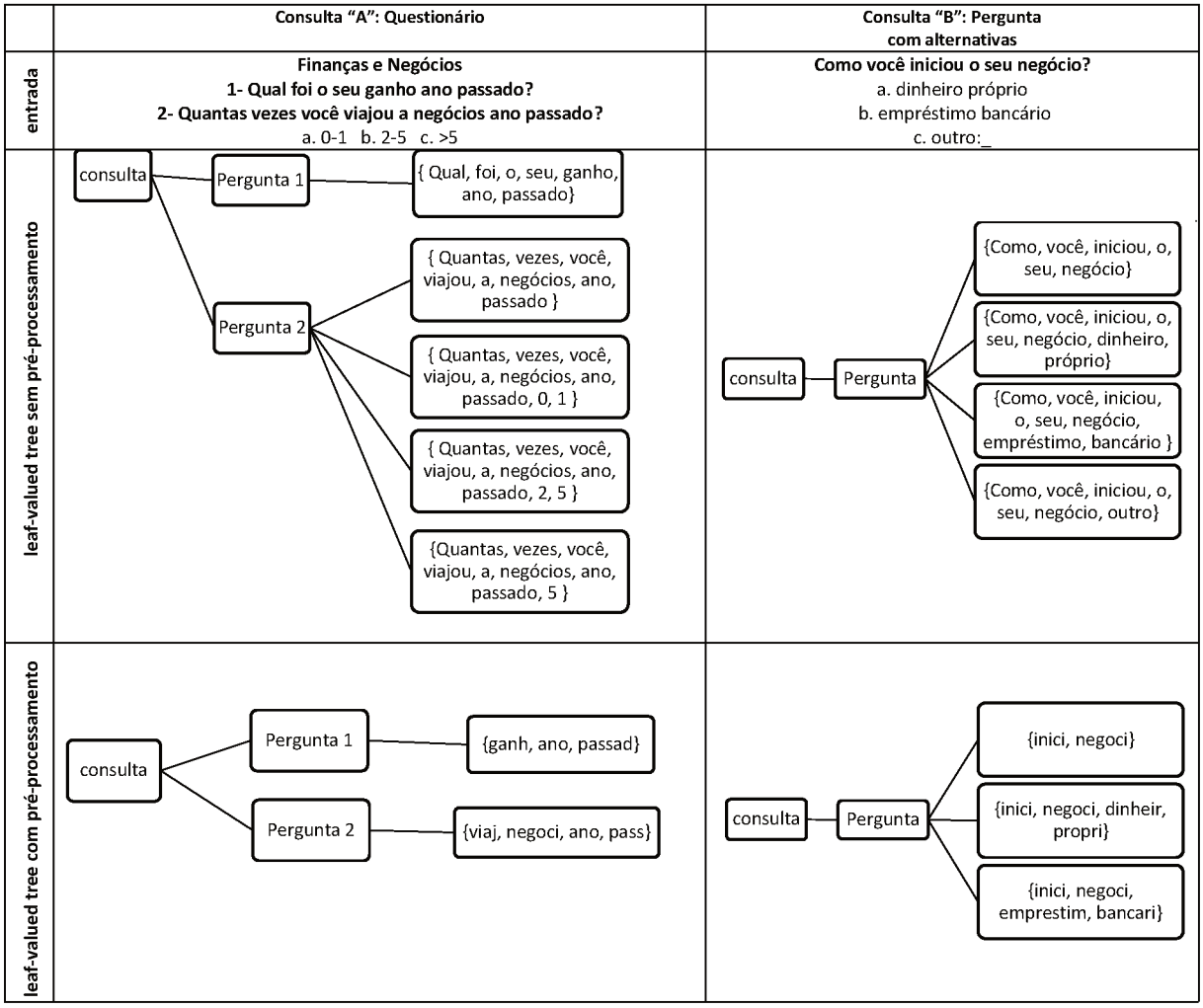

Figura 8 – *Leaf-valued tree*: com pré-processamento vs sem pré-processamento

ou "NÃO". Na amostra, 60,19% das perguntas contêm verbos auxiliares.

Palavras interrogativas. Há um conjunto de palavras que são comumente usadas na formulação de perguntas, tais como "cite", "explique" e "justifique". Na amostra, 10,17% das perguntas contêm esse tipo de palavra.

Outros pronomes. A remoção de outros pronomes tais como "você" e "seu", também melhora a precisão na recuperação dos questionários. Na amostra, 77,31% das perguntas contêm esses pronomes.

O Quadro 12 ilustra um exemplo de duas perguntas e os resultados de um pré-processamento comumente usado e do pré-processamento proposto. Observe que,

nesse exemplo, quando um pré-processamento comumente usado (BAEZA; RIBEIRO-NETO, 2011; MANNING; SCHIITZE, 1999) é aplicado, pode-se notar que há dois *tokens* (ond, voc) que aparecem em ambas as perguntas. Em contraste, no pré-processamento proposto, o número de palavras correspondentes é zero. Isso significa que as perguntas não têm similaridade nas palavras relevantes. Vale ressaltar que, os termos retirados podem ser importantes para uma análise semântica das perguntas, então sugere-se como trabalho futuro utilizar os termos em uma segunda etapa, após a recuperação dos questionários relevantes.

Seguindo a mesma ideia, a remoção de palavras e números comumente utilizados nas alternativas é proposto:

Números. Há alternativas que contêm números, conforme pode ser observado na Figura 7, consulta “A”. Na amostra de 510 questionários, dentre as 5432 alternativas, 1079 são números, ou seja 19,86% das alternativas contêm números.

Escala. A remoção de palavras comumente utilizadas para representar escalas de resposta, por exemplo a escala Likert (LIKERT, 1974), tais como “discordo” e “concordo”. Na amostra, 32,40 % das alternativas contêm essas palavras.

Outras palavras. A remoção das demais palavras comumente usadas nas alternativas, tais como “outro”, “sim” e “não”. Na amostra, 5,64% das alternativas contêm essas palavras.

O pré-processamento das alternativas pode reduzir o número de folhas na *leaf-valued tree*. A Figura 8 ilustra essa redução no número de folhas e também a redução da quantidade de palavras dentro das folhas devido ao pré-processamento como um todo. Nesse exemplo, a consulta “A” reduziu de 5 folhas para apenas 2, já a consulta “B” reduziu de 4 para 3 folhas.

4.5 SIMILARIDADE

Nesta seção, a proposta de métrica de similaridade (QSM - *Questionnaire Similarity Matching*) é descrita. A similaridade entre uma consulta e um questionário é dada pela Definição 11, em que uma média ponderada é calculada para obter um valor de similaridade. O cálculo leva em consideração a estrutura *leaf-valued tree*. A partir dos conjuntos de *tokens* de cada folha da *leaf-valued tree* da consulta do usuário e de um questionário, a ideia é contabilizar o número de *tokens* que são iguais (TES - *Token Equal Score*) entre as folhas e também o número de *tokens* que são sinônimos (TSS - *Token Synonym Score*), calculados respectivamente usando as Definições 1 e 2.

Definição 1 ***Token Equal Score***. *Sejam s_1 e s_2 dois conjuntos de tokens, o valor de*

Token Equal Score é dado pela equação:

$$TES_{(s_1,s_2)} = \frac{\frac{En_{(s_1,s_2)}}{count(s_1)} + \frac{En_{(s_1,s_2)}}{count(s_2)}}{2} \tag{10}$$

a qual retorna uma pontuação entre [0,1], onde En é a $s_1 \cap s_2$ e count contabiliza o número de elementos de cada conjunto.

Definição 2 ***Token Synonym Score****. Seja s_1 e s_2 dois conjuntos de tokens, o valor de Token Synonym Score é dado pela equação:*

$$TSS_{(s_1,s_2)} = \frac{En_{(ss_1,s_2)}}{count(s_2)} \tag{11}$$

a qual retorna uma pontuação entre [0,1], onde ss_1 é um conjunto de tokens que são sinônimos dos tokens do conjunto s_1, En é a $ss_1 \cap s_2$, e count é o número de elementos de conjunto s_2.

Note que a Definição 2 tem uma estrutura diferente da Definição 1, isso ocorre porque o número de sinônimos obtidos de s_1 tende a reduzir o valor a ser obtido se considerar a estrutura da Definição 1.

Para cada nó da *leaf valued tree* (obtida da consulta do usuário) obtêm-se o conjunto de TES e de TSS conforme as Definições 3 e 4.

Definição 3 ***Conjunto de TES****. Sejam p_1 e p_2 dois nós (perguntas) de leaf-valued trees, sendo p_1 do questionário do usuário e p_2 do questionário do dataset, o conjunto de TES é dado por:*

$$\begin{aligned} Ctes_{(p_1,p_2)} = \{&TES(p_{1_1}, p_{2_1}), ..., TES_{(p_{1_1}, p_{2_m})}, \\ &..., \\ &TES_{(p_{1_n}, p_{2_1})}, ..., TES_{(p_{1_n}, p_{2_m})}\} \end{aligned} \tag{12}$$

onde Ctes é o conjunto de valores de TES para cada folha do nó de p_1, p_{1_1} representa a primeira folha do nó p_1 (enunciado da pergunta 1), p_{1_n} é a última folha do nó p_1 (enunciado da pergunta 1 + última opção de resposta da pergunta 1) onde n é o número de folhas do nó p_1, p_{2_1} representa a primeira folha do nó p_2, p_{2_m} é a última folha do nó p_2 onde m é o número de folhas do nó p_2.

Definição 4 ***Conjunto de TSS****. Sejam p_1 e p_2 dois nós (perguntas) de leaf-valued trees, sendo p_1 do questionário do usuário e p_2 do questionário do dataset, o conjunto de TSS é dado por:*

$$\begin{aligned} Ctss_{(p_1,p_2)} = \{&TSS(p_{1_1}, p_{2_1}), ..., TSS_{(p_{1_1}, p_{2_m})}, \\ &..., \\ &TSS_{(p_{1_n}, p_{2_1})}, ..., TSS_{(p_{1_n}, p_{2_m})}\} \end{aligned} \tag{13}$$

onde Ctss é o conjunto de valores de TSS para cada folha do nó de p_1, p_{1_1} representa a primeira folha do nó p_1, p_{1_n} é a última folha do nó p_1 onde n é o número de folhas do nó p_1, p_{2_1} representa a primeira folha do nó p_2, p_{2_m} é a última folha do nó p_2 onde m é o número de folhas do nó p_2.

Para cada nó da *leaf valued tree* da consulta (do usuário) são calculados os valores de *Node Token Equal Score* e *Node Token Synonym Score* conforme as Definições 5 e 6.

Definição 5 ***Node Token Equal Score****. Sejam p_1 e p_2 dois nós (perguntas) de leaf-valued trees, sendo p_1 do questionário do usuário e p_2 do questionário do dataset, a pontuação é dada pela equação:*

$$N_{TES(p_1,p_2)} = MAX(Ctes_{(p_1,p_2)}) \tag{14}$$

N_{TES} é o valor da pontuação entre dois nós de leaf-value trees, e a função MAX retorna o maior valor de token equal score do conjunto Ctes.

Definição 6 ***Node Token Synonym Score****. Sejam p_1 e p_2 dois nós (perguntas) de leaf-valued trees, sendo p_1 do questionário do usuário e p_2 do questionário do dataset, a pontuação é dada pela equação:*

$$N_{TSS(p_1,p_2)} = MAX(Ctss_{(p_1,p_2)}) \tag{15}$$

N_{TSS} é o valor da pontuação entre dois nós de leaf-value trees, e a função MAX retorna o maior valor de token Synonym score do conjunto Ctss.

Note que as Definições 5 e 6 retornam o maior valor (MAX), isso ocorre porque o enunciado da pergunta é repetido quando há opções de resposta, então utilização de retornar o maior valor é para não aumentar o valor de perguntas com alternativas em relação a perguntas sem alternativas.

Após o cálculo de *Node Token Equal Score* e *Node Token Synonym Score* para cada nó da *leaf value tree* da consulta, tem-se o conjunto de valores conforme as Definições 7 e 8.

Definição 7 ***Conjunto de*** *N_{TES}. Sejam U e Q duas leaf-valued trees (questionários), sendo que U é o questionário fornecido pelo usuário e Q um questionário do dataset, e u_1 um nó (pergunta) de U, o conjunto de N_{TES} é dado por:*

$$C_{N_{(u_1,Q)}} = \{N_{TES(u_1,q_1)}, ..., N_{TES(u_1,q_k)}\} \tag{16}$$

onde C_N é o conjunto de valores de N_{TES}, q_1 representa o primeiro nó de Q, e q_k é o ultimo nó de Q, e k representa o número de nós (perguntas) da leaf-valued tree Q.

Definição 8 ***Conjunto de*** N_{TSS}*. Sejam U e Q duas leaf-valued trees (questionários), sendo que U é o questionário fornecido pelo usuário e Q um questionário do dataset, e* u_1 *um nó (pergunta) de U, o conjunto de* N_{TSS} *é dado por:*

$$C_{Ns_{(u_1,Q)}} = \{N_{TSS(u_1,q_1)}, ..., N_{TSS(u_1,q_k)}\} \tag{17}$$

onde C_{Ns} *é o conjunto de valores de* N_{TSS}*,* q_1 *representa o primeiro nó de Q, e* q_k *é o ultimo nó de Q, e k representa o número de nós (perguntas) da leaf-valued tree Q.*

Para cada nó da consulta é calculado o *Highest Token Equal Score* e *Highest Token Synonym Score* conforme as Definições 9 e 10.

Definição 9 ***Highest Token Equal Score****. Sejam U e Q duas leaf-valued trees (questionários), sendo que U é o questionário fornecido pelo usuário e Q um questionário do dataset, e* u_1 *um nó (pergunta) de U, a pontuação é dada pela equação:*

$$H_{TES(u_1,Q)} = MAX(C_{N_{(u_1,Q)}}) \tag{18}$$

Onde H_{TES} *é Highest Token Equal Score de um nó em relação a um questionário e MAX é função que retorna o maior valor do conjunto* C_N*.*

Definição 10 ***Highest Token Synonym Score****. Sejam U e Q duas leaf-valued trees (questionários), sendo que U é o questionário fornecido pelo usuário e Q um questionário do dataset, e* u_1 *um nó (pergunta) de U, a pontuação é dada pela equação:*

$$H_{TSS(u_1,Q)} = MAX(C_{Ns_{(u_1,Q)}}) \tag{19}$$

Onde H_{TSS} *é Highest Token Equal Score de um nó em relação a um questionário e MAX é função que retorna o maior valor do conjunto* C_{Ns}*.*

Note que as Definições 9 e 10 retornam o maior valor (MAX), isso ocorre porque o número de perguntas variam de questionário para questionário, então o maior valor é utilizado para evitar que um questionário que tenha muitas perguntas tem uma relevância maior do que um questionário que tenha um número menor de perguntas.

Considerando que Q_1 é uma *leaf-valued tree* gerada a partir da consulta do usuário e que Q_2 é uma *leaf-valued tree* de um questionário, a similaridade entre Q_1 e Q_2 é calculada conforme a Definição 11.

Definição 11 ***Questionnaire Score****. Sejam* Q_1 *e* Q_2 *duas leaf-valued trees, a similaridade é calculada conforme a equação:*

$$QS_{(Q_1,Q_2)} = \frac{((\frac{1}{n}\sum_{i=1}^{n}(H_{TES}(Q_1,Q_2))_i) * WE) + ((\frac{1}{n}\sum_{i=1}^{n}(H_{TSS}(Q_1,Q_2))_i) * WS)}{WE + WS} \tag{20}$$

a qual retorna uma pontuação entre [0,1], onde i é um nó, H_{TES} é o maior valor de TES de cada folha do nó i, H_{TSS} é o maior valor de TSS para cada folha do nó i, WE é o peso para a pontuação H_{TES}, WS é o peso para a pontuação H_{TSS}, e n é o número de nós da leaf-valued tree gerada da consulta do usuário.

4.6 EXEMPLO

Para exemplicar o cálculo de similaridade, considera-se a Consulta "A" da Figura 7 e o questionário da Figura 9 contendo duas perguntas: 1. Quanto é sua renda para esse ano? 2. Quando você viaja, qual o motivo ? a. Negócios b. Turismo c. Outro. A Figura 9 ilustra a *leaf-valued tree* do questionário. Para cada folha da *leaf-valued tree* da consulta é calculado os valores de TES e TSS em relação a todas as folhas da *leaf-valued tree* do questionário. Para obter os sinônimos utilizou-se a base de dados da openWordnet-pt [3](PAIVA; RADEMAKER; MELO, 2012) .

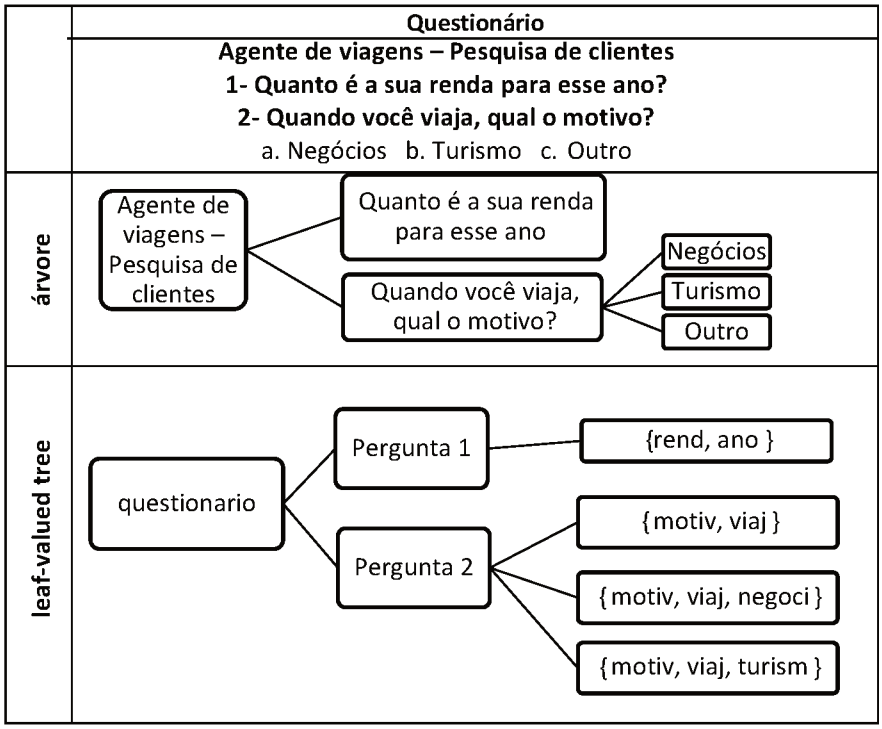

Figura 9 – Questionário

A Tabela 1 mostra o resultado do cálculo de TES e TSS entre as folhas da consulta e do questionário, bem como os valores de H_{TES} e H_{TSS} a serem utilizados na Definição 11. Note que, para a primeira folha da consulta, o único valor maior que zero para TES e TSS é obtido ao calcular tais valores em relação à primeira folha do questionário. Nesse caso, TES é 0,4166 ($\frac{\frac{1}{3}+\frac{1}{2}}{2}$) porque tem um *token* igual ("ano") e TSS é 0,5 ($\frac{1}{2}$) porque as palavras "ganho" e "renda" são consideradas sinonimos[4].

[3] https://github.com/own-pt/openWordnet-PT

[4] https://www.sinonimos.com.br/renda/

Como consequência o valor de H_{TES} para o primeiro nó (Pergunta 1) é 0,4166 e o valor de H_{TSS} é 0,5 já que são os maiores valores de TES e TSS calculados em relação a folha do primeiro nó da consulta. Para o segundo nó da consulta, o valor de H_{TES} é de 0,5833 porque é o maior valor de TES encontrado em relação a folha da consulta com todas as folhas do questionário. Vale ressaltar que o maior valor foi obtido devido a junção da pergunta com a alternativa. Já o valor de H_{TSS} é zero por não haver sinônimos.

Após o cálculo dos valores de TES e TSS, e consequentemente obter os valores de H_{TES} e H_{TSS}, é possível calcular a similaridade conforme a Definição 11. Para esse exemplo, o valor de similaridade é de 0,44996 (QS = ($\frac{(\frac{0,4166+0,5833}{2})*4)+(\frac{0,5+0,0}{2})}{5}$)). Nesse exemplo foi utilizado o valor de 4 para a variável WE e o valor de 1 para WS.

Tabela 1 – Exemplo dos valores de H_{TES} e H_{TSS}

Entrada			**Pontuação das folhas**		**Pontuação dos nós**	
nó da consulta	**folhas da consulta**	**folhas do questionário**	**TES**	**TSS**	**H_{TES}**	**H_{TSS}**
Pergunta 1	ganh, ano, passad	rend, ano	**0,4166**	**0,5**	**0,4166**	**0,5**
		motiv, viaj	0,0	0,0		
		motiv, viaj, negoci	0,0	0,0		
		motiv, viaj, turism	0,0	0,0		
Pergunta 2	viaj, negoci, ano, passad	rend, ano	0,375	0,0	**0,5833**	**0,0**
		motiv, viaj	0,375	0,0		
		motiv, viaj, negoci	**0,5833**	0,0		
		motiv, viaj, turism	0,2916	0,0		

4.7 ALGORITMO DE SIMILARIDADE

O Algoritmo 2 é utilizado para calcular a similaridade entre o questionário fornecido pelo usuário e um questionário do *dataset*. O algoritmo recebe como entrada duas *leaf valued-tree* (linha 1), uma correspondendo ao questionário fornecido pelo usuário (U_{Lvt}) e outra o questionário do *dataset* (Q_{Lvt}). Primeiramente, o algoritmo inicializa as variáveis *H* (H é o conjunto de valores H_{TES} conforme as Definições 9 e 11) e *Hs* (Hs é o conjunto de valores H_{TSS} conforme as Definições 10 e 11) com o conjunto vazio (linhas 2 e 3). Para cada nó da *leaf-valued tree* do usuário, as variáveis s e s_1 são inicializadas com conjunto vazio (linhas 5 e 6). Então, para cada nó da *leaf-value tree* do questionário, é calculado o maior valor de *token equal score* em relação ao nó da *leaf-valued tree* do usuário, e o resultado é agregado ao conjunto de valores de

Algoritmo 2 Questionnaire Score

```
1: Entrada: U_Lvt, Q_Lvt
2: H ← {}
3: Hs ← {}
4: para cada u ∈ U_Lvt faça
5:     s ← {}
6:     s_1 ← {}
7:     para cada q ∈ Q_Lvt faça
8:         s ← s ∪ {N_TES(u, q)}                ▷ Algoritmo 3
9:         s_1 ← s_1 ∪ {N_TSS(u, q)}            ▷ Algoritmo 4
10:    fim para
11:    H ← H ∪{MAX({s})}
12:    Hs ← Hs ∪{MAX({s_1})}
13: fim para
14: S_H ← Σ_{H_TES ∈ H} H_TES / |U|
15: S_Hs ← Σ_{H_TSS ∈ Hs} H_TSS / |U|
16: return ((S_H * WE) + (S_Hs * WS)) / (WE + WS)
```

Algoritmo 3 Maior Token Equal Score

```
1: função N_TES(u, q)
2:     S ← {}
3:     para cada u' ∈ u faça
4:         para cada q' ∈ q faça
5:             S ← S ∪ {TES(u', q')}
6:         fim para
7:     fim para
8:     return MAX(S)
9: fim função
```

token equal score em s (linha 8, ver Algoritmo 3). Da mesma forma, para cada nó da *leaf-value tree* do questionário, é calculado o maior valor de *token synonym score* em relação ao nó da *leaf-valued tree* do usuário, e o resultado é agregado ao conjunto de valores de *token synonym score* em s_1 (linha 9, ver Algoritmo 4). Depois, para cada nó da *leaf-valued tree* do usuário, o maior valor de TES é adicionado ao conjunto de valores de H_{TES} em H. Também para cada nó da *leaf-valued tree* do usuário, o maior valor de TSS é adicionado ao conjunto de valores de H_{TSS} em H_s. A média dos valores contidos em H (linha 14) é calculada, e a média dos valores contidos em H_s (linha 15) também é calculada. Por fim, o valor do *Questionnaire Score* é calculado e retornado.

O Algoritmo 3 é utilizado para retornar o valor de *token equal score* entre dois nós. A função N_{TES} recebe dois nós (linha 1), sendo um referente ao nó de uma *leaf-valued tree* do questionário do usuário e a outra sendo um nó de uma *leaf-valued tree* do questionário do dataset. O valor de TES é calculado para cada folha do nó da *leaf-valued tree* do usuário com cada folha do *leaf-valued tree* (linha 5). Após calcular todos os valores possíveis, o algoritmo retorna apenas o maior valor calculado (linha 8).

Algoritmo 4 Maior Token Synonym Score

```
1: função N_TSS(u, q)
2:     S ← {}
3:     para cada u' ∈ u faça
4:         u_s ← getSynonym(u')
5:         para cada q' ∈ q faça
6:             S ← S ∪ {TSS(u_s, q')}
7:         fim para
8:     fim para
9:     return MAX(S)
10: fim função
```

O Algoritmo 4 é utilizado para retornar o valor de *token synonym score* entre dois nós. A função N_{TSS} recebe dois nós (linha 1), sendo um referente ao nó de uma *leaf-valued tree* do questionário do usuário e a outra sendo um nó de uma *leaf-valued tree* do questionário do dataset. Os *tokens* sinônimos de cada folha da *leaf valued tree* do usuário são armazenados temporariamente em u_s (linha 4). O valor de TSS é calculado para cada folha do nó da *leaf-valued tree* do usuário com cada folha do *leaf-valued tree* (linha 6). Após calcular todos os valores possíveis, o algoritmo retorna apenas o maior valor calculado (linha 9).

4.7.1 Análise de complexidade e de corretude

O Algoritmo 3 mostra o cálculo da similaridade (maior *token equal score*) entre dois nós da *leaf-valued tree*. O *token equal score* é calculado no tempo $O(n^3 log_2 n)$ (n é o número de folhas da *leaf-valued tree* do usuário), já que a execução do cálculo de TES na linha 5 demora $O(nlogn)$, e a passagem por todas as folhas nas linhas 3-9 demora n^2 (O algoritmo termina encontrando o maior elemento do conjunto S que demanda tempo não significativo em uma análise assintótica). Da mesma forma pode-se considerar a mesma análise para o Algoritmo 4.

O Algoritmo 2 mostra o cálculo de similaridade entre dois questionários. A similaridade (QS) é calculada no tempo $O(|U|.|Q|.(n^3 log_2 n))$, já que a execução da função do Algoritmo 3 (F_{TES}) na linha 6 demora $O(n^3 log_2 n)$, e a passagem por todas as perguntas nas linhas 3-9 demora tempo $O(|U|.|Q|)$ ($|U|$ é o número de nós (perguntas) na *leaf-valued tree* do usuário e $|Q|$ é o número de nós na *leaf-valued tree* do questionário do *dataset*).

Lema 1: Quando o Algoritmo 3 para, ele retorna o maior *equal token score* (Definição 5) entre dois nodos u e q de arvores *leaf-valued tree*.

Prova: Para demostrar a corretude do Algoritmo 3, se dividirá a prova em dois passos. Primeiro verifica-se que ele para e depois se ele é correto. O cálculo de TES demanda tempo finito, pois se utiliza intersecções e contagem de elementos em dois conjuntos finitos. Os laços das linhas 3 e 4 repetem um número de vezes igual a cardinalidade

dos conjuntos u e q. Como u e q são conjuntos finitos, conclui-se que o Algoritmo 3 termina. Para concluir a prova, resta demostrar que a linha 8 retornará o maior valor de TES encontrado, o que é trivial e é correto afirmar, pois a função MAX retornará o elemento de maior valor de TES do conjunto S (TES calculado para cada par de u e q, presentes no conjunto S).■

Lema 2: Quando o Algoritmo 4 para, ele retorna o maior *equal synonym score* (Definição 5) entre dois nodos u e q de arvores *leaf-valued tree*.

Prova: Para demostrar a corretude do Algoritmo 4, se dividirá a prova em dois passos. Primeiro verifica-se que ele para e depois se ele é correto. O cálculo de TSS demanda tempo finito, pois se utiliza intersecções e contagem de elementos em dois conjuntos finitos. Os laços das linhas 3 e 5 repetem um número de vezes igual a cardinalidade dos conjuntos u e q. Como u e q são conjuntos finitos, conclui-se que o Algoritmo 4 termina. Para concluir a prova, resta demostrar que a linha 8 retornará o maior valor de TSS encontrado, o que é trivial e é correto afirmar, pois a função MAX retornará o elemento de maior valor de TSS do conjunto S (TSS calculado para cada par de u e q, presentes no conjunto S).■

Teorema: Quando o Algoritmo 2 para, será retornado a similaridade QS (Equação 20) entre os questionários U e Q no formato de *leaf-valued tree*.

Prova: A prova é realizada em dois passos. Primeiro, verifica-se o algoritmo para. Segundo, se quando para retorna o valor de QS para U e Q. Os laços da linha 4 e 7 dependem do número de elementos em U e Q. Como U e Q possuem tamanho finito, então os laços se repetem um número finito de vezes. Na linha 8, o Algoritmo 3 sempre para retornando o maior *equal token score* das perguntas u e q de acordo com o Lema 1. Na linha 9, o Algoritmo 4 sempre para retornando o maior *synonym token score* das perguntas u e q de acordo com o Lema 2. Conclui-se então que o algoritmo para. Para concluir a prova, resta demostrar que a linha 16 retornará o valor referente a média ponderada dos valores de H_{TES} e H_{TSS} encontrados, o que é correto afirmar, pois é realizado uma média ponderada conforme a Equação 20 encontrados nas linhas 11 e 12.■

5 EXPERIMENTOS

Neste capítulo, os principais experimentos realizados são apresentados e analisados. Os experimentos realizados têm dois objetivos principais: (i) determinar a efetividade da busca de questionários; (ii) comparar a abordagem proposta com a de trabalhos existentes.

As seguintes métricas clássicas foram utilizadas para avaliar os resultados: revocação, precisão, *f-value*, DCG (*Discounted Cumulated Gain*), MAP (*Mean Average Precision*) e MAE (*Mean absolute error*) (BAEZA; RIBEIRO-NETO, 2011). Tais métricas são úteis para comparar o desempenho de modelos de busca.

A quantidade de consultas está de acordo com uma regra geral que é frequentemente aplicada no campo de recuperação de informações (BUCKLEY; VOORHEES, 2000). Por essa regra, um conjunto de consultas de tamanho entre 25 e 50 é geralmente empregado na avaliação dos mecanismos de busca.

5.1 METODOLOGIA

A Figura 10 ilustra as etapas realizadas nos experimentos. A primeira etapa foi a coleta dos questionários. A segunda etapa foi a seleção dos parâmetros a serem utilizados na busca e ordenação de questionários. A terceira etapa foi a geração do *ground truth*. Essas três etapas estão detalhadas na próxima Seção 5.2.

A seleção do modelo de busca é realizada na etapa 4. O primeiro modelo selecionado foi o modelo vetorial. A escolha do modelo leva em consideração o mapeamento da literatura e as demais escolhas também levam em consideração a análise dos resultados da etapa 6.

Na etapa 5 é realizada toda a implementação e/ou configuração do modelo escolhido. Nessa etapa são realizadas as buscas conforme as entradas definidas na etapa 2.

A análise dos resultados é realizada na etapa 6. As principais métricas utilizadas são a precisão, revocação e o DCG. Os resultados são armazendos para futura comparação com os demais modelos a serem utilizados. Essa etapa contribui para a abordagem proposta neste trabalho.

A comparação entre os resultados dos modelos de busca é realizada na etapa 7. Nessa etapa utilizam-se as demais métricas (NDCG, MAP e MAE). Com base no *ground truth*, os *rankings* gerados pelos modelos e abordagens foram comparados e avaliados. Nessa etapa procura-se entender porque um modelo é mais eficaz que outro. Por fim (etapa 8), a avaliação dos resultados é realizada e documentada.

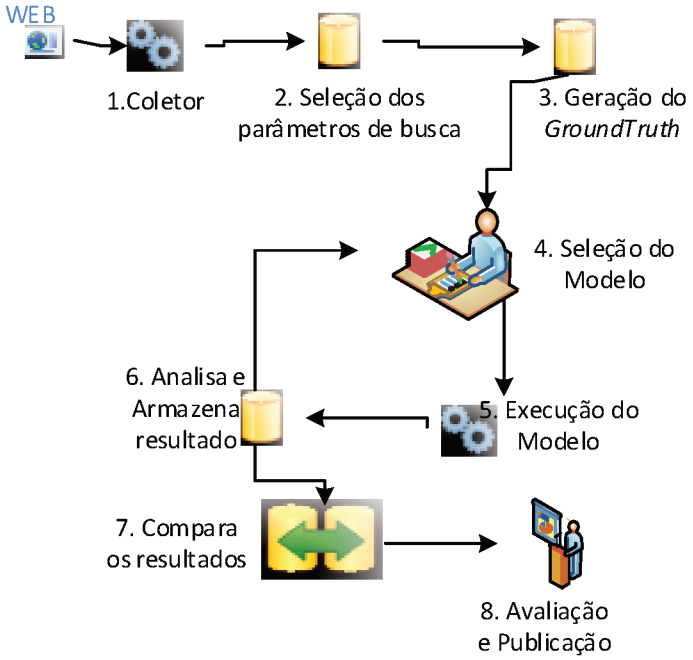

Figura 10 – Processo do experimento

5.2 GROUND TRUTH

Um dos desafios encontrados nesse trabalho foi a criação do *ground truth*, uma vez que não foi encontrado um *dataset* com questionários previamente ordenados. Então a primeira etapa foi coletar questionários na Web. O *dataset* consiste em 510 questionários coletados aleatoriamente da Web por meio de um coletor.

O coletor (*crawler*) foi desenvolvido por um aluno de TCC (Trabalho de Conclusão de Curso) (MATHIAS, 2017). O TCC foi orientado de forma específica para atender a esta tese. Alguns exemplos de questionários de pesquisa serviram de base para contrução do coletor, respeitando o modelo proposto na Seção 4.2. Os questionários coletados são provenientes de diferentes domínios de pesquisa, como mostrado na Tabela 2.

Tabela 2 – Dataset

Domínio	***#qtd***	***Domínio***	***#qtd***
Saúde	86	Tecnologia	40
Marketing	71	Esportes	35
Planejamento de eventos	60	Demografia	32
Comportamento humano	49	Outros	92
Pesquisa com os empregados	45		

Conforme ilustrado na Tabela 2, os questionários coletados são de diferentes domínios de pesquisa, sendo 86 questionários da área da saúde. Já na área de marketing foram coletados 71 questionários. Para planejamento de eventos foram coletados 60 questionários, em relação a pesquisa do comportamento humano foram coletados

49 questionários. Em se tratando de avaliar o ambiente interno das empresas foram coletados 45 questionários. Na área de tecnologia foram coletados 40 questionários de pesquisa. Na área de esportes foram coletados 35 questionários. Sobre demografia foram coletados 32 questionários. Os últimos 92 questionários são de assuntos diversos, como, por exemplo, um questionário de pesquisa com alunos de graduação.

Os questionários coletados contêm, em média, 18 perguntas, sendo que o menor contém 5 perguntas e o maior contém 57. Todos os questionários juntos somam 9.180 perguntas. Deste total, 1.815 são de perguntas abertas, 7.365 são de perguntas fechadas, resulta que 80,22% (7.365) das perguntas coletadas.

Na segunda etapa, após a coleta de dados, foi realizada uma análise dos dados e a seleção de "parâmetros de busca". Segundo Buckley e Voorhees (2000), um conjunto de consultas deve ter um tamanho entre 25 e 50 para ser empregado na avaliação dos mecanismos de busca. Nesse sentido, foram escolhidos 200 questionários que serviram de "parâmetros de busca" para as consultas realizadas nos experimentos. Os "parâmetros de busca" foram separados em quatro categorias de forma a ter 50 consultas para cada categoria: (i) 50 consultas cujo parâmetro de busca é composto por uma palavra (por exemplo, a parte D da Figura 7); (ii) 50 consultas cujo parâmetro de busca é composto por uma sentença (frase ou pergunta aberta, sem alternativas, por exemplo, parte C da Figura 7); (iii) 50 consultas cujo parâmetro de busca é composto por uma pergunta com alternativas (Por exemplo a parte B da Figura 7); (iv) 50 consultas cujo parâmetro de busca é composto por um questionário (Por exemplo, parte A da Figura 7).

O critério para a escolha dos "parâmetros de busca" foi de que existisse no mínimo um questionário similar no *dataset*, ou seja, aleatóriamente foi escolhido um questionário que serviu como fornecedor do parâmetro de busca. Porém, se for utilizado um questionário igual a outro na base de dados, ao fazer a busca, o resultado deve retornar esse questionário como primeiro na ordenação. Então, como o intuito é analisar o comportamento dos mecanismos de busca, optou-se em utilizar apenas parte do questionário, em alguns casos apenas uma palavra, ou frase, ou uma pergunta, ou algumas das perguntas presentes no questionário. Vale ressaltar que, ao analisar a base de questionários em busca do segundo questionário similar, a análise se estendia até encontrar um questionário similar, ou seja, nesta etapa, não foi verificado se existiam mais do que dois questionários similares.

Com os "parâmetros de busca" definidos, a etapa de ordenação dos questionários é realizada para a geração do *ground truth*. Um grupo de três especialistas[1] colaboraram para construir o *ground truth*. A ordenação do *ground truth* foi realizada em 6 fases:

[1] Doutores em educação e com mais de 10 anos experiência em pesquisa e docência no ensino superior.

Tabela 3 – Exemplo de Ground Truth

Consulta	1- Qual foi o seu ganho ano passado? 2- Quantas vezes você viajou a negócios ano passado? a. 0-1 b. 2-5 c. > 5

Ranking	**Questionário**	**Relevância (Sim /Não)**	**Pontuação [0,2] Relevância**
1º	29	Sim	2
2º	70	Sim	2
3º	73	Sim	1
4º	72	Sim	1
5º	79	Sim	1
6º	07	Sim	1
-	1	Não	0
		...	
-	84	Não	0

Fase 1: Ordenação preliminar, para ajudar na elaboração do *ground truth*, foi utilizado o modelo vetorial na ordenação inicial por se tratar de um modelo clássico.

Fase 2: Análise preliminar, após a ordenação da fase 1, foi avaliado se os questionários da posição 50 em diante poderia ser classificado como não relevante (Tabela 3). Então, uma leitura de todas questões de cada questionário, foi realizada para decidir se realmente os questionários que não foram bem classificados não são relevantes.

Fase 3: Ordenação manual, as 200 consultas foram divididas entre os 3 especialistas, então, cada especialista fez a ordenação de forma manual, foram verificados e registrados quais questionários são relevantes, e em que ordem eles devem ser apresentados para cada consulta. Também foi solicitado a indicação da pontuação de relevância, sendo que quanto maior o número, maior seria a relevância.

Fase 4: Revisão, nessa fase cada especialista revisa a ordenação das consultas dos outros 2 especialistas. A ideia é garantir que a ordenação está correta.

Fase 5: Análise das divergências, nessa fase é verificado se houve divergências entre os especialistas e o porquê de cada divergência.

Fase 6: Consolidação, nessa fase é definido a ordenação final, as divergências foram resolvidas após a discussão entre os especialistas e o autor desta tese.

A Tabela 3 contém um exemplo de ordenação de questionários. Nota-se, nesse exemplo, que para consulta apresentada há apenas 6 questionários relevantes, conforme pode ser observado pela coluna “Relevância (Sim / Não)”, a qual indica se o questionário é relevante ou não. Observa-se também que não é necessário “ordenar” os questionários classificados como “não” relevantes. Além da ordenação, foi solicitado que as pessoas indicassem uma escala de relevância. Após a consolidação da ordenação, a pontuação de relevância ficou entre 0 e 2 (coluna “Relevância da Pontuação [0,2]”), onde 2 indica que o questionário é de maior relevância, 1 indica que o questi-

onário é relevante e 0 quando o questionário não é relevante. A escala de relevância foi necessária para poder aplicar a métrica de avaliação NDCG@k (Ganho acumulado com desconto normalizado) nos experimentos.

Quadro 13 – Versões do QSM

Sigla	Descrição
QSM	Abordagem proposta conforme descrito na Seção 4
QSM_{lev}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Levenshtein (Lev).
QSM_{sw}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Smith Waterman (SW).
QSM_{JM}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Jaro Metric (JM).
QSM_{HD}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Hamming Distance (HD).
QSM_{SDM}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Soundex Distance Metric (SDM).
QSM_{CDF}	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade Covington's distance function (CDF).
$QSM_{2\text{-gram}}$	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade *n*-gram com $n = 2$ (2-gram).
$QSM_{3\text{-gram}}$	QSM com a substituição da equação 10 descrita na Definição 1 pelo cálculo de similaridade nd $n = 3$ (3-gram) .
QSM_s	QSM sem o pré-processamento definido na Seção 4.4
QSM_n	QSM sem o pré-processamento definido na Seção 4.4 e sem a parte do cálculo dos sinônimos.

5.3 RESULTADOS

O experimento realizado teve como finalidade realizar a comparação do QSM com outras abordagens (modelos vetorial e *fuzzy*) e verificar qual das versões do QSM é mais eficaz. O Quadro 13 mostra as diferenças das versões do QSM. A abordagem proposta foi refinada de modo a realizar a comparação entre as folhas da estrutura *leaf-value tree*, ou seja, a similaridade é obtida em comparação de *strings* "curtas". Nesse sentido, foram desenvolvidas mais oito versões (QSM_{lev}, QSM_{sw}, QSM_{JM}, QSM_{HD}, QSM_{SDM}, QSM_{CDF}, $QSM_{2\text{-gram}}$ e $QSM_{3\text{-gram}}$) da abordagem para utilizar algoritmos de similaridade em *strings* "curtas" (ver Seção 2.3.1). Para cada versão, foi adaptada a forma de cálculo da similaridade entre as folhas, de modo a utilizar os métodos já conhecidos na literatura com o intuito de verificar qual se adequa melhor na abordagem QSM.

Outra questão que o experimento ajudou a verificar é se a adição do pré-processamento proposto na Seção 4.4, em conjunto com o cálculo com sinônimos,

melhoraria a eficácia da obordagem. Nesse sentido, foram desenvolvidas duas versões (QSM_s e QSM_n). QSM é a versão completa conforme descrito na Seção 4.5.

As Tabelas 4, 5, 6 e 7 apresentam o resultado do cálculo da precisão, revocação, *f-value* [2], MAP e MAE para as quatro categorias de consultas. Os melhores resultados foram destacados, com negrito e os valores sublinhados estão em segundo lugar. O resultado demonstra que a abordagem QSM apresenta, na média, resultados superiores em contrapartida ao uso dos modelos vetorial e *fuzzy* ao se tratar de recuperação de questionários, independentemente do tipo de categoria da consulta.

Tabela 4 – Resultado das consultas da categoria *i* (uma palavra)

Similaridade	**Precisao**	**Revocação**	***F-value***	**MAP**	**MAE**
QSM	**0,90168**	**0,939648**	**0,921868**	**0,882745**	**0,487357**
QSM_{lev}	0,896547	0,938885	0,917687	0,878616	0,491178
QSM_{sw}	0,8778	0,94956	0,912271	0,83391	0,641343
QSM_{JM}	0,892848	0,90924	0,900969	0,86517	0,709077
QSM_{HD}	0,881216	0,939816	0,909573	0,845967	0,568781
QSM_{SDM}	0,833256	0,94644	0,886249	0,814091	0,669601
QSM_{CDF}	0,777084	0,91056	0,838544	0,767759	0,709077
$QSM_{2\text{-}gram}$	0,8624	0,934336	0,896928	0,836528	0,51049
$QSM_{3\text{-}gram}$	0,830247	0,887689	0,858008	0,812812	0,531648
QSM_s	0,75376	0,93744	0,835625	0,722856	0,562178
QSM_n	0,666064	0,969248	0,789551	0,658737	0,671092
Vetorial	0,64584	0,882246	0,745756	0,625819	0,903276
Fuzzy	0,653475	0,895668	0,84148	0,661736	0,862934

Tabela 5 – Resultado das consultas da categoria *ii* (uma sentença)

Similaridade	**Precisao**	**Revocação**	***F-value***	**MAP**	**MAE**
QSM	**0,762065**	**0,8065523**	**0,783678**	**0,738441**	**0,57232**
QSM_{lev}	0,674454	0,7613178	0,715258	0,66029	0,71367
QSM_{sw}	0,74613	0,772854	0,73543	0,715539	0,59239
QSM_{JM}	0,758921	0,807126	0,775824	0,729466	0,58155
QSM_{HD}	0,708268	0,804474	0,753312	0,68702	0,77898
QSM_{SDM}	0,749034	0,7988436	0,733137	0,725814	0,74504
QSM_{CDF}	0,660521	0,771976	0,712762	0,634101	0,79896
$QSM_{2\text{-}gram}$	0,73304	0,7941856	0,7623891	0,724977	0,60082
$QSM_{3\text{-}gram}$	0,70571	0,7545357	0,729306	0,697241	0,59385
QSM_s	0,640696	0,796824	0,710282	0,627882	0,73725
QSM_n	0,566154	0,7238608	0,671118	0,543508	0,84778
Vetorial	0,586092	0,7944508	0,674548	0,568509	1,28492
Fuzzy	0,548964	0,7499091	0,633893	0,521516	1,32629

A precisão do modelo *fuzzy* é um pouco melhor que a do modelo vetorial na categoria de consulta *i*, mas nas categorias *ii*, *iii* e *iv* o modelo vetorial é melhor. De

[2] Considerando os 20 questionários melhores classificados (ordenados)

Tabela 6 – Resultado das consultas da categoria *iii* (uma pergunta)

Similaridade	**Precisao**	**Revocação**	***F-value***	**MAP**	**MAE**
QSM	**0,690341**	**0,723683**	**0,70669**	**0,66894**	**1,03342**
QSM_{lev}	0,607008	0,685186	0,64373	0,59426	1,31169
QSM_{sw}	0,671517	0,6955686	0,68922	0,64398	1,28961
QSM_{JM}	0,683029	0,7264134	0,69788	0,66731	1,10365
QSM_{HD}	0,674133	0,7189592	0,69582	0,65323	1,25622
QSM_{SDM}	0,637441	0,7240266	0,67798	0,61831	1,10855
QSM_{CDF}	0,594469	0,6965784	0,64148	0,57069	1,12853
$QSM_{2\text{-}gram}$	0,659736	0,714767	0,68615	0,65247	1,1818
$QSM_{3\text{-}gram}$	0,635139	0,6790821	0,65637	0,62751	1,15514
QSM_{s}	0,576626	0,565745	0,57113	0,56509	1,04917
QSM_{n}	0,509539	0,5849412	0,54464	0,48915	1,21001
Vetorial	0,391546	0,5640601	0,45518	0,37012	1,65204
Fuzzy	0,389764	0,5324355	0,45006	0,37027	1,70523

Tabela 7 – Resultado das consultas da categoria *iv* (questionário)

Similaridade	**Precisao**	**Revocação**	***F-value***	**MAP**	**MAE**
QSM	**0,621307**	**0,6513147**	**0,63854**	**0,5927**	**1,2918**
QSM_{lev}	0,546307	0,6166674	0,57188	0,5382	1,4931
QSM_{sw}	0,604365	0,6337721	0,62544	0,5795	1,3952
QSM_{JM}	0,614726	0,6470633	0,62827	0,5823	1,3718
QSM_{HD}	0,606717	0,6260117	0,61523	0,5835	1,3903
QSM_{SDM}	0,573697	0,6516239	0,61731	0,5505	1,4719
QSM_{CDF}	0,535022	0,6269206	0,58194	0,5268	1,4801
$QSM_{2\text{-}gram}$	0,593762	0,6432903	0,61547	0,5724	1,4254
$QSM_{3\text{-}gram}$	0,571625	0,6111739	0,58899	0,5626	1,5987
QSM_{s}	0,518964	0,6077774	0,55889	0,5026	1,6587
QSM_{n}	0,458585	0,5789241	0,51647	0,4411	1,7488
Vetorial	0,289784	0,6149787	0,38894	0,2863	1,8356
Fuzzy	0,207564	0,6213454	0,35615	0,20607	1,8947

forma que, quanto maior for o número de palavras na consulta, mas eficaz é o modelo vetorial.

Quando apenas uma palavra é usada como parâmetro de pesquisa (Tabela 4), a precisão é aproximadamente 25% menor para as versões QSM_{s} e QSM_{n} (versões sem pré-processamento). A versão sem pré-processamento, mas com sinônimos, obteve uma precisão aproximadamente 10% maior do que a sem contabilizar os sinônimos. A diferença é mínima entre as demais versões do QSM. Destaque para as versões QSM, QSM_{lev}, QSM_{JM} e QSM_{HD} que têm praticamente a mesma precisão com QSM em primeiro e QSM_{lev} em segundo. Considerando a métrica *f-value*, as versões QSM, QSM_{lev} e QSM_{sw} se destacam com a melhor relação precisão e revocação. O mesmo ocorre com a qualidade da revocação, métrica MAP, que registraram menor erro abso-

luto (MAE).

Em consultas cuja entrada é uma sentença (Tabela 5), há uma pequena diferença entre as versões do QSM, sendo que QSM tem a melhor precisão e o QSM_{JM} fica em segundo. Já as versões sem pré-processamento, a precisão é aproximadamente 12% menor do que a do melhor colocado. Considerando a métrica *f-value* as versões QSM, QSM_{JM} e QSM_{HD} se destacam com a melhor relação precisão e revocação. O mesmo ocorre com a qualidade da revocação, métrica MAP, e registraram menor erro absoluto (MAE).

Nas consultas de categoria *iii* (pergunta com alternativas) e *iv* (questionário) (Tabelas 6 e 7), o destaque fica para as versões QSM, QSM_{JM}, QSM_{HD} e QSM_{sw},

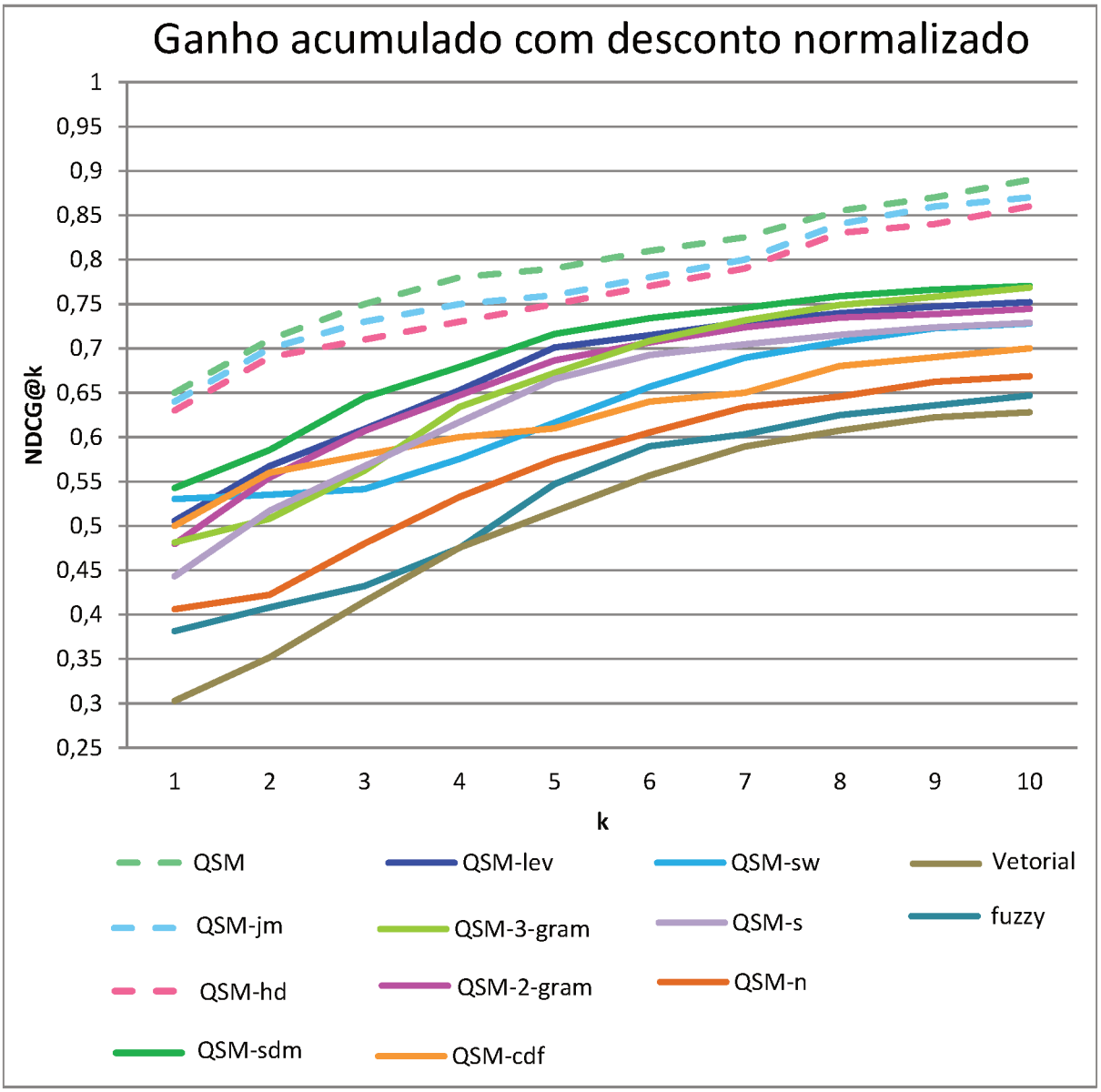

Figura 11 – NDCG

cuja a diferença é mínima. A diferença para os modelos sem o pré-processamento proposto praticamente se mantém. E a diferença da precisão em relação aos modelos vetorial e *fuzzy* aumenta para praticamente 40%. O QSM obteve o menor erro absoluto (MAE), destacam-se também QSM_{JM} e QSM_{HD}. QSM, QSM_{JM} e QSM_{HD} também se destacam com a melhor relação precisão e revocação (*f-value*). O mesmo ocorre com a qualidade da revocação, métrica MAP.

Observou-se que o principal motivo do QSM ser mais eficaz do que as outras oito versões (QSM_{lev}, QSM_{sw}, QSM_{JM}, QSM_{HD}, QSM_{SDM}, QSM_{CDF}, $QSM_{2\text{-}gram}$ e $QSM_{3\text{-}gram}$) é que as outras oito versões pontuam questionários não relevantes. Por exemplo, no algoritmo de *Levenshtein*, considere as palavras morre e corre, a distância é de 1 alteração, pontuando uma pergunta que não tem pontuação no QSM.

A Figura 11 mostra o gráfico gerado pelo cálculo do NDCG@10 (Ganho acumulado com desconto normalizado para as dez primeiras posições [3]). É possível ver que o QSM, QSM_{JM}, QSM_{HD} tiveram um NDCG@10 mais alto em 100% das consultas do que as outras.

O resultado da análise do experimento revelou que a abordagem QSM foi mais eficaz que os demais modelos. As versões QSM_{JM}e QSM_{HD} também se destacam. De maneira geral, todas as 8 versões do QSM (que utilizam o pré-processamento) têm um desempenho satisfatório. Já as versões do QSM sem o pré-processamento e os modelos vetorial e *fuzzy* ficaram abaixo dos demais.

5.3.1 Considerações Finais

Foram realizados outros experimentos durante a evolução do presente trabalho. O primeiro foi a realização do experimento apenas com o modelo vetorial, com intuito de descobrir como se comportava a ordenação dos questionários utilizando um modelo clássico. O resultado foi publicado em (SOUZA; DORNELES, 2017a).

Outros experimentos foram realizados e publicados em (SOUZA; DORNELES, 2017b, 2018, 2019b). Os resultados demonstram a eficácia da abordagem proposta em relação a outros modelos de busca de documentos, tais como o modelo vetorial. Para a execução dos testes preliminares, utilizou-se a ideia de contagem de *tokens* iguais (Definição 5). Após avaliação dos resultados, a abordagem proposta foi refinada de modo a realizar a comparação entre as folhas da estrutura *leaf-value tree*, ou seja, a similaridade é obtida em comparação de *strings* “curtas”.

A publicação no evento SBSI2018 (SOUZA; DORNELES, 2018) [4] demonstram os experimentos que comparam a abordagem QSM e o modelo vetorial em relação a recuperação de questionários. O trabalho publicado no SBSI2018 foi estendido em:

[3] K na Figura 11 são as posições; por exemplo, apenas k = 1 refere-se aos questionários que retornaram primeiro. Se k = 10, então é a precisão dos 10 primeiros

[4] O trabalho foi considerado um dos 5 melhores do evento: XIV Simpósio Brasileiro de Sistemas de informação, https://www.ucs.br/site/eventos/sbsi2018/

(SOUZA; DORNELES, 2019a). O trabalho apresenta a comparação das abordagens QSM, vetorial e *fuzzy*. Os resultados foram semelhantes aos do experimento descrito em (SOUZA; DORNELES, 2018). A diferença principal foi a adição do modelo *fuzzy* na comparação, onde o *fuzzy* foi um pouco melhor que o modelo vetorial e o QSM apresentou melhor performance em relação aos dois modelos testados.

A publicação (SOUZA; DORNELES, 2019b) no DocEng2019 [5] compara o QSM com as seguintes baselines: (i) duas métricas de similaridade para árvore, TED e GPD, propostas respectivamente em (YANG; KALNIS; TUNG, 2005; PAWLIK; AUGSTEN, 2012); (ii) o algoritmo que combina o modelo vetorial com métricas de similaridade para árvore, proposto em (CHIM; DENG, 2007) e é chamado de NSTC; (iii) o modelo vetorial (VM). Os experimentos mostram que o QSM é mais eficaz na recuperação de questionários do que outros sistemas em todas as consultas. Com base nos resultados, pode-se afirmar que combinar *tokens* de pergunta e possíveis *tokens* de resposta, e manter a árvore como uma estrutura hierárquica, são características importantes a serem levadas em consideração na comparação de questionários de pesquisa. O sistema de pré-processamento aumentou significativamente a eficácia da recuperação dos questionários.

Por fim, foi submetido um artigo para um periódico com os experimentos das *strings* curtas (resultado apresentado na Seção 5.3). Os experimentos mostram que *Jaro Metric*, *Hamming Distance* e *Token Equal Score* no QSM são mais eficazes na recuperação de questionários do que outros métodos em todas as consultas. Após a análise dos resultados, pode-se afirmar que a estratégia descrita no QSM é eficaz na ordenação de questionários.

[5] The 19th ACM Symposium on Document Engineering, http://doceng.org/doceng2019/

6 CONCLUSÃO E TRABALHOS FUTUROS

A recuperação de questionários se mostrou desafiadora. As abordagens clássicas como o modelo vetorial e *fuzzy* apresentaram resultados não satisfatórios na ordenação dos questionários. A falta de um *dataset* para execução de experimentos demonstra a necessidade de extração de questionários e o desenvolvimento de uma *baseline* para execução de experimentos. Portanto, uma das contribuições do presente trabalho é o *ground truth* gerado para ser utilizado nos experimentos, e também possibilita trabalhos futuros.

Outro ponto é a estrutura do questionário, o qual, se for tratado como um documento de texto não apresenta bons resultados na ordenação. Mas, se o questionário for representado em estrutura de árvore, a precisão nos resultados da ordenação aumenta consideravelmente. Partindo desse princípio, o presente trabalho apresenta uma estrutura auxiliar, chamada *leaf-value tree*, que converte o questionário em uma estrutura de árvore, de modo a aumentar a precisão do cálculo de similaridade entre questionários, sendo, portanto, mais uma contribuição do presente trabalho.

O cálculo da similaridade, levando em consideração as opções de resposta da pergunta se mostrou eficaz. A verificação de sinônimos é outro fator que ajudou a melhorar o grau de similaridade entre questionários e por consequência melhorar a precisão e a qualidade da ordenação dos questionários relevantes. De modo que, a principal contribuição do presente trabalho é o cálculo de similaridade entre questionários.

Após a execução dos experimentos, observou-se que a abordagem proposta é eficaz na busca por similaridade entre questionários de pesquisa. Portanto, concluise que o objetivo geral do presente trabalho foi alcançado. Assim como os objetivos específicos do presente trabalho, uma vez que um modelo conceitual para recuperação de questionários foi descrito e regras de similaridade foram apresentadas e testadas na execução do algoritmo para o cálculo da similaridade entre questionários. Também foram definidos quatro formatos para realização de consultas.

Como trabalhos futuros, os seguintes pontos devem ser investigados:

1. Indexação de questionários: Desenvolvimento de uma plataforma que realize a extração de questionários da Web, faça a indexação questionários e ordenação dos questionários relevantes e também exiba as respostas. Desse modo, sugere-se como trabalho futuro a questão de como indexar questionários? e de como agilizar a busca? Em conjunto pode-se realizar uma pesquisa de trabalhos que verificam a qualidade e ordenação das respostas, uma vez que a recuperação de respostas não faz parte do escopo do presente trabalho.

2. Parametrização da medida de similaridade: Como adicionar medidas ou parâme-

tros que alteram um pouco a similaridade de acordo com o interesse do usuário? adição de pesos nos operadores de consulta a questionários que podem ser compostos. Uma motivação poderia ser o *Google Forms* e outras ferramentas do tipo, que oferece o recurso de fazer um formulário a partir de um *template*. Um mecanismo automatizado poderia trazer *templates* com mais significado ao usuário.

3. Uso da métrica proposta ao contexto de CQA: Como misturar questionários e CQA no mesmo modelo? Por que não? Utilizar os tópicos (assuntos) como guia para agrupar as perguntas?

4. Uso de *embeddings* para incrementar a métrica de similaridade. Uma possibilidade é a troca do uso de sinônimos por *word embeddings*, onde teria uma conjunto de palavras associadas, por exemplo pizza e sanduiche estão associadas a comida.

5. Verificar e ajustar a similaridade de acordo com a área do questionário. A ordenação de questionários foi tratada de forma genérica, isto é, independente da área de assunto do questionário. Então é interessante realizar experimentos em cada área. Como sugestão de trabalho futuro, por exemplo, na área da medicina onde existe um vocabulário controlado, poderia ter as seguintes perguntas de pesquisa: será que o uso de sinônimos é mais ou menos eficaz? ou a troca de pesos no uso de sinônimos para diferentes áreas de pesquisa resultaria em melhores resultados?

6. A utilização de técnicas de *data mining* é também uma sugestão de trabalho futuro. Uma interessante questão de pesquisa seria: o uso de aprendizagem de máquina melhoraria a eficiência e a qualidade na recuperação de questionários?

Todas essas questões levantadas são interessante e, embora estejam fora do escopo do presente trabalho, servem de motivação para a continuação da pesquisa na área de recuperação de questionários.

REFERÊNCIAS

ABBASI, Ahmed; CHEN, Hsinchun; SALEM, Arab. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. **ACM Transactions on Information Systems (TOIS)**, ACM, v. 26, n. 3, p. 12, 2008.

ABDELHAMID, Neda; AYESH, Aladdin; THABTAH, Fadi. Phishing detection based associative classification data mining. **Expert Systems with Applications**, Elsevier, v. 41, n. 13, p. 5948–5959, 2014.

ABDUL-MAGEED, Muhammad; DIAB, Mona; KÜBLER, Sandra. SAMAR: Subjectivity and sentiment analysis for Arabic social media. **Computer Speech & Language**, Elsevier, v. 28, n. 1, p. 20–37, 2014.

AGICHTEIN, Eugene *et al.* Finding high-quality content in social media. *In*: ACM. PROCEEDINGS of the 2008 international conference on web search and data mining. [*S.l.*: *s.n.*], 2008. P. 183–194.

AHMED, Nizar A *et al.* Scalable multi-label arabic text classification. *In*: IEEE. INFORMATION and Communication Systems (ICICS), 2015 6th International Conference on. [*S.l.*: *s.n.*], 2015. P. 212–217.

ANDRADE, Maria Margarida de *et al.* **Introdução à metodologia do trabalho científico**. [*S.l.*]: São Paulo: Atlas, 1999.

ANWAR, Tarique; ABULAISH, Muhammad. Ranking radically influential web forum users. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 10, n. 6, p. 1289–1298, 2015.

APACHE Lucene. [*S.l.*: *s.n.*], 2020. https://lucene.apache.org/. (Accessed: 01-2020).

ARTINO, Anthony R *et al.* Developing questionnaires for educational research: AMEE Guide No. 87. **Medical Teacher**, v. 36, n. 6, p. 463–474, jun. 2014.

ASLAM, Javed A; YILMAZ, Emine. A geometric interpretation and analysis of R-precision. *In*: PROCEEDINGS of the 14th ACM international conference on Information and knowledge management. [*S.l.*: *s.n.*], 2005. P. 664–671.

ASLAY, Çiğdem *et al.* Competition-based networks for expert finding. *In*: ACM. PROCEEDINGS of the 36th international ACM SIGIR conference on Research and development in information retrieval. [*S.l.*: *s.n.*], 2013. P. 1033–1036.

BAEZA, YR; RIBEIRO-NETO, Berthier. **Modern Information Retrieval-the concepts and technology behind search**. [*S.l.*]: Pearson, 2011.

BILENKO, Mikhail *et al.* Adaptive name matching in information integration. **IEEE Intelligent Systems**, IEEE, v. 18, n. 5, p. 16–23, 2003.

BIOLCHINI, Jorge *et al.* Systematic review in software engineering. **System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES**, v. 679, n. 05, p. 45, 2005.

BIYANI, Prakhar *et al.* Using non-lexical features for identifying factual and opinionative threads in online forums. **Knowledge-Based Systems**, Elsevier, v. 69, p. 170–178, 2014.

BOYNTON, Petra M; GREENHALGH, Trisha. Selecting, designing, and developing your questionnaire. **BMJ : British Medical Journal**, v. 328, n. 7451, p. 1312–1315, mai. 2004.

BUCKLEY, Chris; VOORHEES, Ellen M. Evaluating evaluation measure stability. *In*: ACM. PROCEEDINGS of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. [*S.l.*: *s.n.*], 2000. P. 33–40.

CARVALHO, ACPLF *et al.* Inteligência Artificial–uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, 2011.

CHAPMAN, Sam. String similarity metrics for information integration. **http://www. dcs. shef. ac. uk/˜ sam/stringmetrics. html**, 2007.

CHEN, Danqi *et al.* Reading wikipedia to answer open-domain questions. **arXiv preprint arXiv:1704.00051**, 2017.

CHEN, Ruey-Cheng *et al.* Harnessing semantics for answer sentence retrieval. *In*: ACM. PROCEEDINGS of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval. [*S.l.*: *s.n.*], 2015. P. 21–27.

CHIM, Hung; DENG, Xiaotie. A new suffix tree similarity measure for document clustering. *In*: ACM. PROCEEDINGS of the 16th international conference on World Wide Web. [*S.l.*: *s.n.*], 2007. P. 121–130.

CILIBRASI, Rudi L; VITANYI, Paul MB. The google similarity distance. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 19, n. 3, 2007.

CLAUDINO, Lucas Paravizo; NUNES, Murilo Barbosa; SILVA, FC da. Finanças pessoais: um estudo de caso com servidores públicos. **Anais do SEMEAD-Seminários em Administração, São Paulo, SP, Brasil**, v. 12, 2009.

COELHO, Maria Teresa Vieira. **Comunicação terapêutica em Enfermagem: utilização pelos enfermeiros**. 2015. Tese (Doutorado) – Instituto de Ciências Biomédicas Abel Salazar.

DRINGUS, Laurie P; ELLIS, Timothy. Using data mining as a strategy for assessing asynchronous discussion forums. **Computers & Education**, Elsevier, v. 45, n. 1, p. 141–160, 2005.

FAULKNER, Adam Robert. Automated classification of argument stance in student essays: A linguistically motivated approach with an application for supporting argument summarization, 2014.

FERRARI, DANIEL GOMES; SILVA, LEANDRO NUNES DE CASTRO. **Introdução a mineração de dados**. [*S.l.*]: Saraiva Educação SA, 2017.

FERREIRA, Cirillo Ribeiro. Classi cação não-supervisionada hierárquica de artigos jornalísticos, 2015.

FOSTER JR, Blair; EVANS, Dr Patricia. Comparative q-gram Analysis of Gene Promoter Regions. **New Brunswick**, 2003.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. [*S.l.*]: Gulf Professional Publishing, 2005.

GRAPPY, Arnaud *et al.* Selecting answers to questions from Web documents by a robust validation process. *In*: IEEE. WEB Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. [*S.l.*: *s.n.*], 2011. P. 55–62.

GU, Jiatao; WANG, Changhan; ZHAO, Junbo. Levenshtein transformer. *In*: ADVANCES in Neural Information Processing Systems. [*S.l.*: *s.n.*], 2019. P. 11181–11191.

GUIMARÃES, Lucas Marques Sathler; MEIRELES, Magali Rezende Gouvêa; ALMEIDA, Paulo Eduardo Maciel de. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. **Perspectivas em Ciência da Informação**, v. 24, n. 1, p. 169–190, 2019.

GUPTA, Sparsh; CARVALHO, Vitor R. FAQ retrieval using attentive matching. *In*: PROCEEDINGS of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. [*S.l.*: *s.n.*], 2019. P. 929–932.

HALL, Patrick AV; DOWLING, Geoff R. Approximate string matching. **ACM computing surveys (CSUR)**, ACM, v. 12, n. 4, p. 381–402, 1980.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. [*S.l.*]: Elsevier, 2011.

HOVY, Eduard H *et al.* Question Answering in Webclopedia. *In*: TREC. [*S.l.*: *s.n.*], 2000. P. 53–56.

HU, Fanghuai; RUAN, Tong; SHAO, Zhiqing. Complete-thread extraction from web forums. *In*: SPRINGER. ASIA-PACIFIC Web Conference. [*S.l.*: *s.n.*], 2012. P. 727–734.

HUANG, Jonathan *et al.* Superposter behavior in MOOC forums. *In*: ACM. PROCEEDINGS of the first ACM conference on Learning@ scale conference. [*S.l.*: *s.n.*], 2014. P. 117–126.

KANNAN, Subbu; GURUSAMY, Vairaprakash. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2014.

KIM, Su Nam; CAVEDON, Lawrence; BALDWIN, Timothy. Classifying dialogue acts in one-on-one live chats. *In*: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2010 Conference on Empirical Methods in Natural Language Processing. [*S.l.*: *s.n.*], 2010. P. 862–871.

KIM, Sung-min; HA, Young-guk. Automated discovery of small business domain knowledge using web crawling and data mining. *In*: IEEE. BIG Data and Smart Computing (BigComp), 2016 International Conference on. [*S.l.*: *s.n.*], 2016. P. 481–484.

KNUPFER, Nancy Nelson; MCLELLAN, Hilary. **41. DESCRIPTIVE RESEARCH METHODOLOGIES**. [*S.l.*: *s.n.*], 1996.

KONDRAK, Grzegorz. Phonetic alignment and similarity. **Computers and the Humanities**, Springer, v. 37, n. 3, p. 273–291, 2003.

KOTHARI, Chakravanti Rajagopalachari. **Research methodology: Methods and techniques**. [*S.l.*]: New Age International, 2004.

KOWALSKI, Gerald J; MAYBURY, Mark T. **Information storage and retrieval systems: theory and implementation**. [*S.l.*]: Springer Science & Business Media, 2006. v. 8.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos da metodologia científica**. [*S.l.*]: Altas, 2010.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. *In*: 8. SOVIET physics doklady. [*S.l.*: *s.n.*], 1966. P. 707–710.

LIKERT, Rensis. A method of constructing an attitude scale. **Scaling: A sourcebook for behavioral scientists**, Aldine Publishing, Chicago, p. 233–243, 1974.

LIM, Wee Yong; SACHAN, Amit; THING, Vrizlynn LL. A lightweight algorithm for automated forum information processing. *In*: IEEE COMPUTER SOCIETY. PROCEEDINGS of the 2013 IEEE/WIC/ACM International Joint Conferences on Web

Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. [*S.l.*: *s.n.*], 2013. P. 121–126.

MALHOTRA, Naresh K. **Pesquisa de Marketing-: Uma Orientação Aplicada**. [*S.l.*]: Bookman Editora, 2001.

MANNING, Christopher D; SCHIITZE, Hinrich. **Foundations of Statistical Natural Language Processing**. [*S.l.*]: MIT press, 1999.

MANNING, Christopher D; SCHÜTZE, Hinrich *et al.* **Foundations of statistical natural language processing**. [*S.l.*]: MIT Press, 1999. v. 999.

MATHIAS, Gilney Nathanael. qFex: um crawler para busca e extração de questionários de pesquisa em documentos HTML. Florianópolis, SC., 2017.

MEUSEL, Robert; MIKA, Peter; BLANCO, Roi. Focused crawling for structured data. *In*: ACM. PROCEEDINGS of the 23rd ACM International Conference on Conference on Information and Knowledge Management. [*S.l.*: *s.n.*], 2014. P. 1039–1048.

MIKOLOV, Tomas *et al.* Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MOLLÉRI, Jefferson Seide; PETERSEN, Kai; MENDES, Emilia. Survey Guidelines in Software Engineering: An Annotated Review. *In*: PROCEEDINGS of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. New York, NY, USA: ACM, 2016. 58:1–58:6.

NAVARRO, Gonzalo. A guided tour to approximate string matching. **ACM computing surveys (CSUR)**, ACM, v. 33, n. 1, p. 31–88, 2001.

ORTIGOSA, Alvaro; MARTÍN, José M; CARRO, Rosa M. Sentiment analysis in Facebook and its application to e-learning. **Computers in Human Behavior**, Elsevier, v. 31, p. 527–541, 2014.

PAIVA, Valeria de; RADEMAKER, Alexandre; MELO, Gerard de. **Openwordnet-pt: An open brazilian wordnet for reasoning**. [*S.l.*], 2012.

PAWLIK, Mateusz; AUGSTEN, Nikolaus. A Robust Algorithm for the Tree Edit Distance, 2012.

PICARD, Claude François. **Graphs and questionnaires**. [*S.l.*]: Elsevier, 1980. v. 32.

REFAEE, Eshrag; RIESER, Verena. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. *In*: LREC. [*S.l.*: *s.n.*], 2014. P. 2268–2273.

ROSENTHAL, Sara *et al.* Semeval-2015 task 10: Sentiment analysis in twitter. *In*: PROCEEDINGS of the 9th international workshop on semantic evaluation (SemEval 2015). [*S.l.*: *s.n.*], 2015. P. 451–463.

RUSSELL, Stuart J; NORVIG, Peter. Artificial intelligence: a modern approach (International Edition). {Pearson US Imports & PHIPEs}, 2002.

SADILEK, Adam *et al.* Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. *In*: CITESEER. AAAI. [*S.l.*: *s.n.*], 2016. P. 3982–3990.

SCHMIDT, Lena; WEEDS, Julie; HIGGINS, Julian. Data mining in clinical trial text: Transformers for classification and question answering tasks. **arXiv preprint arXiv:2001.11268**, 2020.

SHEATSLEY, Paul B *et al.* Questionnaire construction and item writing. **Handbook of survey research**, p. 195–230, 1983.

SMAILOVIC, Jasmina *et al.* Stream-based active learning for sentiment analysis in the financial domain. **Information Sciences**, Elsevier, v. 285, p. 181–203, 2014.

SMITH, Temple F; WATERMAN, Michael S *et al.* Identification of common molecular subsequences. **Journal of molecular biology**, Elsevier Science, v. 147, n. 1, p. 195–197, 1981.

SOUZA MINAYO, Maria Cecília de. **Pesquisa social: teoria, método e criatividade**. [*S.l.*]: Editora Vozes Limitada, 2011.

SOUZA, Richard Henrique de; DORNELES, Carina Friedrich. Analisando a eficácia do modelo vetorial de busca na ordenação de questionários. **XIII Simpósio Brasileiro de Sistemas de Informação**, SBSI 2017, Lavras, MG, BR, jun. 2017.

SOUZA, Richard Henrique de; DORNELES, Carina Friedrich. Comparando a eficácia na recuperação de questionários: QSMatching vs Vector model vs Fuzzy. **iSys-Brazilian Journal of Information Systems**, v. 12, n. 1, p. 100–118, 2019.

SOUZA, Richard Henrique de; DORNELES, Carina Friedrich. QSMatching vs Vector model: comparing effectiveness in questionnaires retrieval. **XIV Simpósio Brasileiro de Sistemas de Informação**, SBSI 2018, Caxias do Sul, RS, BR, jun. 2018.

SOUZA, Richard Henrique de; DORNELES, Carina Friedrich. QSMatching: an approach to calculate similarity between questionnaires. *In*: ACM. PROCEEDINGS of the 19th International Conference on Information Integration and Web-based Applications & Services. [*S.l.*: *s.n.*], 2017. P. 141–145.

SOUZA, Richard Henrique de; DORNELES, Carina Friedrich. Searching and ranking questionnaires: an approach to calculate similarity between questionnaires. *In*: ACM.

PROCEEDINGS of the 19th ACM Symposium on Document Engineering, DocEng2019. [*S.l.*: *s.n.*], 2019.

SRBA, Ivan; BIELIKOVA, Maria. A Comprehensive Survey and Classification of Approaches for Community Question Answering. **ACM Trans. Web**, ACM, New York, NY, USA, v. 10, n. 3, 18:1–18:63, ago. 2016. ISSN 1559-1131. DOI: 10.1145/2934687. Disponível em: http://doi.acm.org/10.1145/2934687.

SRIVIDHYA, V; ANITHA, R. Evaluating preprocessing techniques in text categorization. **International journal of computer science and application**, v. 47, n. 11, p. 49–51, 2010.

VIEIRA, Sonia. **Como elaborar questionários**. [*S.l.*]: Atlas, 2009.

WAMBSGANSS, Thiemo *et al.* A Conversational Agent to Improve Response Quality in Course Evaluations. *In*: EXTENDED Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. [*S.l.*: *s.n.*], 2020. P. 1–9.

WASLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier, 2014. ISBN 9788535277821.

WEN, Miaomiao; YANG, Diyi; ROSE, Carolyn. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *In*: EDUCATIONAL Data Mining 2014. [*S.l.*: *s.n.*], 2014.

WITTEN, Ian; MILNE, David. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *In*: PROCEEDING of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA. [*S.l.*: *s.n.*], 2008. P. 25–30.

YANG, Diyi; PIERGALLINI, Mario *et al.* Forum thread recommendation for massive open online courses. *In*: EDUCATIONAL Data Mining 2014. [*S.l.*: *s.n.*], 2014.

YANG, Rui; KALNIS, Panos; TUNG, Anthony KH. Similarity evaluation on tree-structured data. *In*: ACM. PROCEEDINGS of the 2005 ACM SIGMOD international conference on Management of data. [*S.l.*: *s.n.*], 2005. P. 754–765.