Analisando a eficácia do modelo vetorial de busca na ordenação de questionários

Alternative Title: Analyzing the vector model effectiveness in ordering questionnaires

Richard Henrique de Souza
Universidade Federal de Santa Catarina
Departamento de Informática e Estatística (INE)
Programa de Pós Graduação em Ciência da Computação (PPGCC)
Florianópolis - Santa Catarina (Brasil)
richard.henrique@ufsc.br

Carina Friedrich Dorneles
Universidade Federal de Santa Catarina
Departamento de Informática e Estatística (INE)
Programa de Pós Graduação em Ciência da Computação (PPGCC)
Florianópolis - Santa Catarina (Brasil)
dorneles@inf.ufsc.br

RESUMO

A elaboração de questionários para posterior aplicação em entrevistas, sejam de levantamento estatístico ou para pesquisa científica, não é uma tarefa trivial, pois questões mal elaboradas podem conduzir a respostas diretas e interpretações ingênuas ou sem sentido. Devido a isso, pode ser interessante reaproveitar, parcial ou totalmente, questionários já construídos com a mesma finalidade. Neste artigo, é descrito um experimento para analisar a utilização do modelo vetorial na recuperação de questionários já existentes na Web. Neste sentido, realizou-se um experimento para verificar a eficácia do modelo vetorial na busca e ordenação de questionários. O resultado da análise do experimento revelou que o modelo vetorial foi eficaz na ordenação do primeiro questionário relevante na maioria das consultas realizadas. Contudo, houve uma variação entre os resultados obtidos e os esperados na ordenação dos questionários que deveriam aparecer a partir da segunda posição em diante.

Palavras-Chave

Questionário, modelo vetorial, ordenação.

ABSTRACT

The elaboration of questionnaires for application in interviews, statistical surveys or scientific research is not a trivial task, because poorly worked questions can lead to direct answers with meaningless or naive interpretations. Therefore, it may be interesting to reuse, partially or totally, questionnaires already constructed with the same purpose. In this paper, we describe an experiment to analyze the execution of the vector model in the retrieval of questionnaires. In this sense, an experiment was carried out to verify the vector model effectiveness in searching and ordering questionnaires. The result of the analysis of the experiment revealed that the vector model was effective in ordering the first relevant questionnaire in most of the queries. However, there was a variation between actual and expected results from the ordering of the relevant questionnaires, considering the questionnaires that should appear in the second position.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Information systems** → **Information retrieval; Retrieval models and ranking;** *Evaluation of retrieval results; Presentation of retrieval results;*

Keywords

Questionnaire, vector model, ranking.

1. INTRODUÇÃO

Pesquisa científica pode ser realizada de várias formas, sendo que, a pesquisa descritiva é caracterizada pelo levantamento de dados e pela aplicação de entrevistas e questionários [13]. Além disso, a pesquisa descritiva pode ser considerada um passo prévio para encontrar fenômenos não explicados pelas teorias vigentes [15, 25]. Desse modo, elaborar um questionário útil, representa uma tarefa importante para a pesquisa descritiva. Construir questionários não é uma tarefa fácil, mas aplicar algum tempo e esforço na sua construção pode ser um fator favorável no que se deseja investigar. O questionário deve ser muito bem organizado e conter uma ordem lógica para o entrevistado, evitando uma estrutura confusa e complexa, ou perguntas demasiadas longas [24]. As perguntas de um questionário podem ser classificadas como abertas, fechadas ou de múltipla escolha [14].

Devido aos cuidados e dificuldades concepção de questionários, pode ser interessante reutilizar questionários já existentes, de forma parcial ou total. É possível que muitos questionários semelhantes entre si já tenham sido elaborados e, neste caso, tanto a sua estrutura quanto as respostas poderiam ser reutilizadas. Considerando o ponto de vista do pesquisador, podem-se listar algumas vantagens em se reutilizar questionários similares: (1) as perguntas de questioná-

rios já existentes poderiam ser incorporadas ao questionário de quem está fazendo um novo, com 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 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.

Neste sentido, pode ser interessante a utilização de ferramentas de busca a fim de encontrar questionários similares, tendo em vista que a capacidade de encontrar determinado conteúdo relevante, em um intervalo de tempo aceitável, torna possível a pesquisadores realizarem seus procedimentos de pesquisa o mais breve possível [4, 16]. Embora, até a presente data, não tenha sido encontrado 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)* [22]. Alguns exemplos são: o trabalho de Sung-min and Young-guk [11], Chen *et al* [5], Yang *et al* [26], que recuperam informação em fóruns analisando as perguntas e as respostas dadas pelos usuários dos fóruns. No entanto, 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 ou de múltipla escolha, conforme pode ser observado na Figura 1.

Neste artigo, é apresentada uma proposta de aplicação do algoritmo de ordenação proposto no modelo vetorial para recuperar questionários de pesquisa. No intuito de avaliar a aplicabilidade do modelo sobre questionários de pesquisa é realizado um experimento sobre um conjunto de questionários reais extraídos da Web. Ao aplicar métricas, tais como precisão e a revocação, notou-se que o modelo vetorial, na maioria das consultas, fez a ordenação dos questionários de forma que o primeiro questionário ordenado pelo modelo era o questionário esperado, porém, a partir do segundo, o modelo não se comportou da forma esperada, ou seja, na segunda posição de ordenação em diante os questionários não eram os esperados para aquelas posições.

O restante do artigo está organizado da seguinte forma: na seção seguinte, são descritos alguns conceitos que norteiam o presente trabalho. Na Seção 3, é definido uma proposta de um modelo conceitual abstrato para representar o questionário. Na Seção 4, é realizada uma discussão de como calcular a similaridade entre questionários. O experimento, os resultados, e metodologia adotada são apresentados na Seção 5. A Seção 6 apresenta os trabalhos relacionados. Finalmente, na Seção 7, são mostradas as conclusões e trabalhos futuros.

2. BACKGROUND

Nesta seção, são mostrados alguns conceitos básicos que norteiam esse trabalho. Inicialmente, é dada uma visão geral sobre questionários, e posteriormente uma descrição breve sobre o modelo vetorial, considerando minimamente os conceitos utilizados na proposta apresentada neste artigo.

2.1 Questionário de Pesquisa

Um questionário de pesquisa é um instrumento de coleta de dados, constituído por uma série ordenada de perguntas, respondidas sem a presença do entrevistador, onde as perguntas variam de acordo com as circunstâncias ou com o tipo de investigação [20].

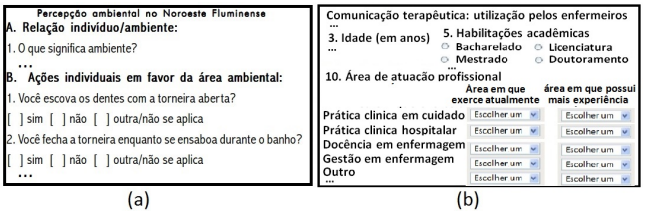

Figure 1: Exemplos de questionários em pesquisas descritivas: (a) Três perguntas retiradas do questionário publicado no artigo de Villar *et al* [24], (b) Três perguntas retiradas do questionário publicado na tese Comunicação terapêutica em Enfermagem: utilização pelos enfermeiros [7]

Questionários podem ser a fonte de obtenção de dados estatísticos que ajudam a realizar estudos de caso, comparativos, argumentações, coletar opiniões que podem ser evidências da validade dos resultados de uma pesquisa [25, 15]. 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, ou o isolamento de variáveis principais [15].

Quanto à forma, as perguntas do questionário, em geral, podem ser de diversos tipos [20, 10, 2, 15, 8, 23, 25], todavia, o presente trabalho considera a classificação das perguntas em tipos: aberto, fechado e múltipla escolha [20, 15]. Perguntas abertas admitem uma quantidade maior de respostas diferentes entre os entrevistados, onde, cada entrevistado pode responder livremente, permitindo que se obtenham mais respostas não coincidentes entre os entrevistados. Nas perguntas fechadas, o pesquisador define as alternativas que podem ser apontadas pelo entrevistado, que deve assinalar aquela que mais se ajusta às suas características, ideias ou sentimentos. Já nas de múltipla escolha são apresentadas várias alternativas e o pesquisado pode assinalar mais de uma delas (respostas múltiplas), podendo ou não, ter uma alternativa aberta que permita uma resposta diferente das alternativas apresentadas [20] .

2.2 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. Então, antes de indexar é necessário realizar o pré-processamento dos documentos (normalização, *stopwords*, substantivos ...) [3].

Um dos cálculos realizados na aplicação do modelo vetorial é o do TF-IDF, usado para definição de pesos nos termos, onde TF é 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 [3] .

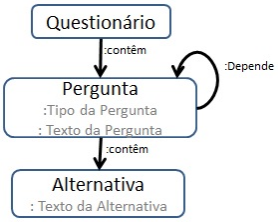

Figure 2: Modelo conceitual de questionário

3. MODELO DE REPRESENTAÇÃO

Nesta seção, é descrita a proposta de um modelo conceitual de representação de questionários de pesquisa, apresentado na Figura 2.

O modelo representa um questionário que é composto por um conjunto de perguntas. Cada pergunta é associada a um tipo de pergunta (aberta, fechada e de múltipla escolha). Uma pergunta pode conter um conjunto de perguntas (ver a pergunta 10 do questionário b da Figura 1). A pergunta poderá conter alternativas, se e somente se, o tipo da pergunta for fechada ou de múltipla escolha. Cada alternativa representa uma opção de resposta para o entrevistado.

Conforme o modelo da Figura 2 e de acordo com a definição de questionário por Picard [18], um questionário pode ser representado por um grafo conectado. Partindo desse princípio, propõe-se um grafo simplificado, da seguinte forma: Q = (V, T), onde V representa os vértices, de tal modo que o conjunto de vértices (V) tem a partição $V = P \cup A$, onde P consiste dos vértices chamados de perguntas e A consiste dos vértices chamados de alternativas. Sendo que, T representa as arestas, onde $T \in (P \times A) \cup (P \times P)$.

4. SIMILARIDADE ENTRE QUESTIONÁRIOS DE PESQUISA

Nesta seção, são descritos os critérios de similaridade entre o parâmetro de busca e os questionários de pesquisa de uma base de documentos, um cenário exemplo que embasa os critérios definidos e o algoritmo de ordenação dos questionários resultantes de uma busca.

4.1 Cenário exemplo

Supondo que um pesquisador construa os parâmetros de busca conforme descrito na Figura 3, com o intuito de averiguar se existem questionários na área temática de finanças, mas especificamente em finanças pessoais (um exemplo de uso de questionários de pesquisa na área temática de finanças pode ser visto em Claudino *et al* [6]). Sendo que os parâmetros de buscas são utilizados para realizar a consulta na base de questionários e considerando a existência de um questionário, conforme pode ser observado na Figura 3. Então, para verificar se o questionário da base é similar ao parâmetro de busca, é a realizada uma análise.

Para facilitar a análise, nesta seção, os parâmetros da consulta mostrados na Figura 3 são chamados apenas como consulta e o questionário de pesquisa da base de questionários é chamado apenas de questionário. Propõe-se fazer uma análise pergunta por pergunta, de forma a procurar quais perguntas da consulta são iguais ou semelhantes, em comparação com as perguntas do questionário. As comparações são efetuadas considerando as seguintes premissas:

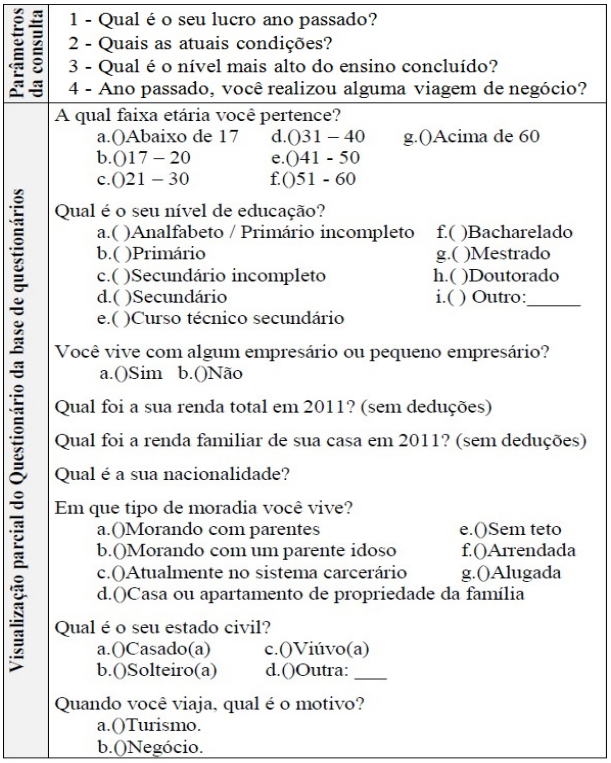

Figure 3: Exemplo de consulta

Premissa 1: uma pergunta p1 é considerada equivalente a uma pergunta p2, se a informação a ser obtida por p1 é igual ou análoga a informação a ser obtida de p2.

Premissa 2: uma pergunta p1 é considerada semelhante a pergunta p2, se a informação a ser obtida por p1 é semelhante ou parcialmente parecida com a informação a ser obtida de p2.

Definidas as premissas, é realizada a comparação das perguntas da consulta com as perguntas do questionário.

Pergunta 1 (Qual é o seu lucro ano passado?) da consulta: Ao realizar uma comparação da Pergunta 1 com as perguntas do questionário, deve-se levar em consideração que a informação a ser obtida na Pergunta 1 da consulta (conforme Premissas 1 e 2) é o ganho que alguém (ou empresa) obteve, isso é importante para verificar se as perguntas são semelhantes, no caso das perguntas não serem escritas rigorosamente da mesma maneira. Assim, é necessário verificar se existe alguma pergunta do questionário que seja igual a Pergunta 1 ou cujo objetivo possa ser considerado semelhante (Premissas 1 e 2). Sendo assim, pode-se dizer 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)" do questionário buscam informações semelhantes. Uma pergunta do questionário busca pela informação do ganho individual e a outra busca a informação do ganho familiar. Além dessa diferença entre as perguntas, tem-se o fato de que a Pergunta 1 da consulta refere-se ao ganho líquido e as duas perguntas do questionário, aparentemente, referem-se ao ganho bruto. Assim, considerando a Premissa 2, a pergunta "Qual é o seu lucro ano passado" da consulta é dita semelhante a duas perguntas do questionário.

Pergunta 2 (Quais as atuais condições?) da consulta: Ao

realizar uma comparação da Pergunta 2 com as perguntas do questionário, observa-se que a informação a ser obtida é muito vaga, podendo ter diversas interpretações. Mas se for considerado o contexto do questionário, uma interpretação possível é que a pergunta se refira a obter a informação das condições financeiras do entrevistado. Mesmo assim se for adotada essa interpretação, a pergunta continua sendo muito vaga, então, ao se comparar a Pergunta 2 com as perguntas do questionário, não se encontra semelhança. Portanto, não existe no questionário, alguma pergunta que seja semelhante ou igual a Pergunta 2 da consulta.

Pergunta 3 (Qual é o nível mais alto do ensino concluído?) da consulta: Ao realizar uma comparação da Pergunta 3 com as perguntas do questionário, observa-se que a informação a ser obtida é o nível mais alto de escolaridade que o entrevistado possui. No questionário, existe a pergunta "Qual é o seu nível de educação?" com alternativas, que também indica que a informação a ser obtida é o nível mais alto de escolaridade que o entrevistado possui. Então, considerando a Premissa 1, a Pergunta 3 é equivalente a uma pergunta do questionário.

Pergunta 4 (Ano passado, você realizou alguma viagem de negócio?) da consulta: Ao realizar uma comparação da Pergunta 4 com as perguntas do questionário, observa-se que a informação a ser obtida é se o entrevistado fez uma viagem de negócio no ano anterior a que foi perguntado. 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. Tal conclusão é possível ao ser considerada a pergunta em conjunto com suas alternativas. Então, considerando a Premissa 2, a Pergunta 4 da consulta é semelhante a uma pergunta do questionário.

Neste cenário, após a comparação das 4 perguntas da consulta com as 9 perguntas do questionário, obteve-se como resultado: 1 pergunta equivalente (pergunta 4 da consulta), 3 perguntas semelhantes (a pergunta 1 da consulta é semelhante a 2 perguntas do questionário que somando a pergunta semelhante encontrada com a pergunta 4 obteve-se o valor 3) e 1 pergunta ausente no questionário. Vale ressaltar que, a consulta e o questionário pode-se verificar a escolaridade, a renda e se fez viagem de negócio.

4.2 Noção de Similaridade

Para que o experimento com o modelo vetorial possa ser avaliado, primeiramente é necessário descobrir qual deveria ser o resultado esperado. Neste sentido, é necessário estabelecer o *ground truth*, ou seja, para cada consulta, estabelecer a lista ordenada (por ordem de relevância) de questionários. Então, para obter as listas ordenadas de questionários, é preciso descobrir o quanto cada questionário é relevante para uma determinada consulta. Neste caso, um questionário é dito mais relevante, se e somente se, o questionário for o mais similar com os parâmetros da consulta. O cálculo da similaridade proposto é dividido em três funções distintas: função de equivalência, função de semelhança e função de ausência.

- **Função de equivalência**: retorna o número de perguntas que o questionário contém que são consideradas equivalentes às perguntas passadas como parâmetro da consulta. Neste contexto, equivalente significa que a pergunta do parâmetro da consulta, na prática, é considerada de mesmo valor da pergunta do questionário.
- **Função de semelhança**: retorna o número de perguntas que o questionário contém que são consideradas semelhantes às perguntas passadas como parâmetro da consulta. Neste contexto, semelhante significa que a pergunta do parâmetro da consulta, na prática, é considerada análoga, de mesma natureza ou de mesmo propósito da pergunta do questionário.
- **Função de ausência**: retorna o número de perguntas do parâmetro da busca que não foram contabilizados pelas funções de equivalência e de semelhança. Neste contexto, diferente das funções anteriores, optou-se por contabilizar as perguntas da consulta que não constam no questionário, para não penalizar questionários que são constituídos de um número maior de perguntas.

O grau de similaridade é calculado a partir do resultado dado pelas 3 funções (equivalência, semelhança e ausência). Com os resultados das funções, o cálculo de similaridade segue as seguintes regras:

- **Regra 1**: Quanto maior for o número de perguntas equivalentes, maior é o grau de similaridade entre o parâmetro de consulta e o questionário, independentemente do resultado das outras duas funções.
- **Regra 2**: Aplica-se a Regra 2, se e somente se, houver empate ao aplicar a Regra 1. Neste caso, quanto maior for o número de perguntas semelhantes, maior é o grau de similaridade entre o parâmetro de consulta e o questionário, restrito aos questionários que obtiveram o mesmo número de perguntas equivalentes em relação ao parâmetro de consulta.
- **Regra 3**: Aplica-se a Regra 3, se e somente se, o número de perguntas equivalentes ou semelhantes forem maior que zero e quando houver empate ao serem aplicadas as Regras 1 e 2. Neste caso, quanto menor for o número de perguntas ausentes, maior é o grau de similaridade entre o parâmetro de consulta e o questionário, restrito aos questionários que obtiveram o mesmo número de perguntas equivalentes e semelhantes em relação ao parâmetro de consulta.
- **Regra 4**: Aplica-se a Regra 4, se e somente se, os resultados das funções de equivalência e de semelhança forem iguais a zero. Neste caso, os questionários serão considerados irrelevantes e descartados da lista ordenada para a consulta cujas as funções de equivalência e de semelhança forem iguais a zero.

Desta forma, um questionário é considerado relevante, se e somente se, o questionário obter um valor maior que zero ao aplicar a função de equivalência ou ao aplicar a função de semelhança, caso contrário aplica-se a Regra 4. A ordenação dos questionários segue as Regras 1, 2 e 3. As funções e regras descritas nesta seção foram desenvolvidas após a análise do cenário descrito na seção 4.1.

4.3 Algoritmo de ordenação

O algoritmo 1 tem como propósito encontrar questionários por ordem de relevância. O algoritmo recebe c como parâmetro para a busca e ordenação, onde c pode ser uma palavra, frase, pergunta ou questionário (ver exemplos na Figura 4). Para cada questionário da base, as variáveis e,s,a

Algorithm 1 Ranking de questionários

```
1: função BuscaQuestionarios(c)
2:     para cada questionário da base faça
3:         e, s, a ← 0
4:         para cada pergunta do questionário faça
5:             se tipopergunta == aberta então
6:                 verifica(pergunta, c)
7:             senão
8:                 para cada alternativa da pergunta faça
9:                     verifica((pergunta ∪ pergunta), c)
10:                fim para
11:            fim se
12:        fim para
13:        se (e > 0)ou(s > 0) então
14:            ordenaQuestionario(questionario, e, s, a)
15:        fim se
16:    fim para
17: fim função
18: função VERIFIVA(p,c)
19:    e ← e + funoEquivalencia(p, c)
20:    s ← s + funoSemelhana(p, c)
21:    a ← a + funoAusencia(p, c)
22: fim função
```

Consulta	Parâmetro de busca
1	Viagem de negócio
2	Qual é o motivo de sua viagem?
3	Qual das seguintes opções melhor descreve a razão para você fazer o curso? Maior exigência Menor exigência Curso eletivo mais avançado Educação geral Curso universitário eletivo
4	Quanto tempo você trabalha na empresa? Qual a sua posição na empresa? Qual o departamento que você trabalha?

Figure 4: Parâmetros de consulta

são inicializadas com zero, sendo que a variável e representa o número de perguntas equivalentes, a variável s representa o número de perguntas semelhantes e a variável a representa o número de perguntas ausentes.

Para cada pergunta do questionário, é verificado se a pergunta é do tipo aberta, caso positivo, é verificada se a pergunta é equivalente ou semelhante ao que foi passado como parâmetro de busca. Caso a pergunta seja fechada ou múltipla escolha, então é necessário verificar a equivalência ou semelhança da pergunta juntamente com suas respectivas alternativas. Se depois de verificar todas as perguntas, o questionário for relevante, ou seja, tem ao menos uma questão equivalente ou uma questão semelhante com o parâmetro de consulta, é realizada a inserção do questionário na lista ordenada de questionários levando em consideração o número de perguntas equivalentes, semelhantes e ausentes. Caso o questionário não seja relevante, o mesmo não é adicionado na lista.

O procedimento *verifica*, executa as funções de equivalência, semelhança e ausência descritas na seção 4.2

5. AVALIAÇÃO EXPERIMENTAL

Os experimentos realizados têm, basicamente, dois principais objetivos: (i) verificar como o modelo vetorial se comporta ao ser usado para realizar a consulta de questionários; e (ii) verificar a efetividade dos critérios de similaridade descritos na Seção 4.2. Na presente seção, são apresentadas as características da base de dados usada, uma descrição breve das métricas utilizadas para avaliação, a metodologia da avaliaçao dos experimentos, e os resultados obtidos juntamente com uma análise do comportamento do processo de busca.

5.1 Dados utilizados e métricas de avaliação

Montar a base de dados com os questionários representa o primeiro passo para a realização do experimento. Portanto, foram coletados 99 questionários disponíveis na web por meio de um *crawler*, os quais foram armazenados em um banco de dados, em uma tabela contendo uma coluna para o texto do questionário e uma coluna para o *link* do qual foi coletado o questionário e chave primária artificial. O texto salvo nesta tabela contém apenas o questionário em si, outros elementos da página web de onde foi extraído o questionário foram retirados do texto, tais como elementos HTML.

Os questionários coletados são de diferentes domínios de pesquisa, sendo que 18 questionários são referentes a pesquisa de satisfação de produtos, clientes, serviços dentre outros. Já na área da saúde foram coletados 15 questionários de pesquisa, sendo que dois são de saúde animal. Sobre pesquisa de mercado foram coletados 10 questionários diferentes tipos de negócio. Para planejamento de eventos foram coletados 8 questionários, mesma quantidade de questionários coletados que tratam de pesquisa de comportamento humano. Em se tratando de avaliar o ambiente interno das empresas foram coletados 7 questionários. Na área de tecnologia foram coletados 6 questionários de pesquisa. Na área de esportes foram coletados 3 questionários. Sobre demografia foram coletados 2 questionários. Os últimos 15 questionários são de assuntos diversos, por exemplo, o questionário de pesquisa com alunos de graduação.

Os questionários coletados contêm em média 17 perguntas, sendo que o menor questionário contêm 5 perguntas e o maior questionário contêm 57 perguntas. Todos os questionários juntos correspondem a 1742 perguntas. Do total de perguntas coletadas, 340 são de perguntas abertas, 1255 são de perguntas fechadas e 147 são de perguntas de múltipla escolha. Verifica-se que 80,48% (1402) das perguntas coletadas contêm alternativas (perguntas fechadas juntamente com as de múltipla escolha).

Na avaliação realizadas, são utilizadas as seguintes métricas clássicas para avaliação dos resultados: revocação, precisão, *f-value*, MAP e DCG (detalhes podem ser encontrados em [3]). As métricas de avaliação são aplicadas para averiguar o desempenho do modelo vetorial em 50 consultas.

5.2 Metodologia adotada

Os experimentos foram conduzidos conforme ilustra a Figura 5. Na Etapa 1, foi realizada a coleta de questionários disponíveis na internet, por meio de um *crawler*. O objetivo é ter uma base de questionário para a execução do experimento. Como resultado desta etapa, foram coletados 99 questionários. O detalhamento da base de dados encontra-se na seção 5.1 .

Na Etapa 2, após a coleta de dados, foi realizada uma aná-

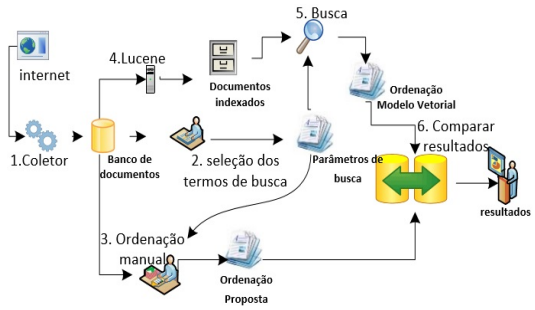

Figure 5: Processo do experimento

lise dos dados. Foram escolhidos 50 questionários que serviram de base para as consultas realizadas no experimento, ou seja, para servir como parâmetro de busca. O critério para a escolha dos parâmetros de busca foi de que existisse no mínimo 2 questionários similares na base, ou seja, o questionário que serviu como fornecedor do parâmetro de busca e mais um questionário. Porém, se utilizar um questionário igual a outro na base de dados ao fazer a consulta, o resultado deve retornar esse questionário como primeiro na ordenação. Então, como o intuito é analisar o comportamento do modelo vetorial, optou-se em utilizar apenas parte do questionário, em alguns casos apenas uma palavra, ou frase, ou uma pergunta, ou algumas das perguntas presente no questionário, conforme os exemplos mostrados na Figura 4. Vale ressaltar que, ao analisar a base de questionários em busca do segundo questionário similar, a análise se estendia até encontrar um questionário, o qual, pelo menos uma pergunta pudesse ser contabilizada conforme as Regras 1 ou 2 descritas na seção 4.2 , ou seja, nesta etapa, não foi verificado se existia mais do que 2 questionários similares.

Na Etapa 3, com intuito de haver uma base de comparação contendo quais questionários deveriam ser retornados das consultas propostas, é realizado o processo de ordenação manual utilizando as funções e regras da seção 4.2. Como resultado, obteve-se uma lista ordenada de questionários relevantes para cada parâmetro de busca, a qual é denominada de *ground truth*. O *ground truth* é utilizado para realizar a comparação com os resultados do modelo vetorial, permitindo assim realizar a análise dos resultados.

Na Etapa 4, os questionários são indexados pelo *Lucene* [1], a qual foi selecionada por ser possível utilizar o modelo vetorial e também por conseguir indexar documentos, uma vez que os questionários podem ser tratados como documentos. Porém, o cálculo de relevância padrão do *Lucene* é baseado em técnicas que combinam o modelo vetorial e booleano, então é necessário configurar a ferramenta para ser utilizado apenas o modelo vetorial.

Na Etapa 5, é realizado o processo de ordenação dos questionários relevantes utilizando o modelo vetorial. Os parâmetros de busca foram selecionados conforme descrito na Etapa 2. O *Lucene* se encarrega de fazer o pré-processamento parâmetros de consulta da mesma forma que foi realizado ao indexar os questionários [1]. Vale destacar, que os parâmetros de consulta podem consistir de um questionário ou uma sentença ou apenas uma palavra.

Na Etapa 6, são comparados os resultados da ordenação manual (*ground truth*) e a ordenação gerada pelo modelo vetorial com o intuito de verificar seu desempenho por meio das métricas: revocação, precisão, *f-value*, MAP e DCG.

5.3 Resultados obtidos

De acordo com o presente experimento, pode-se verificar que o modelo vetorial obteve 88% de eficácia na ordenação do primeiro questionário relevante, ou seja, 44 das 50 consultas executadas, obtive-se uma precisão de 100 % para o primeiro questionário recuperado. Vale ressaltar que a ordenação realizada pelo modelo vetorial obteve uma precisão 100% para os 4 primeiros questionários recuperados em 74% das consultas realizadas, resultando uma alta eficácia para os primeiros 4 pontos de revocação, conforme pode ser observado na Figura 7.

Entretanto, a partir do quarto ponto de revocação, a precisão diminui consideravelmente, como pode ser observado na Figura 7, indicando que questionários não relevantes estão sendo ordenados na frente de questionários relevantes. A Figura 8, mostra o resultado do cálculo das médias da precisão e revocação, além do cálculo de *f-value* considerando os 20 questionários melhores classificados (ordenados) nas 50 consultas realizadas. Como resultado, observa-se que a revocação no modelo vetorial é melhor que a precisão.

A Figura 6 mostra o gráfico gerado ao calcular o DCG (*Discounted Cumulated Gain*) para o modelo vetorial e para o *ground truth* para cada uma das 50 consultas. A escala de relevância adotada é de 0 a 2, onde 2 indica que o questionário possui uma ou mais perguntas equivalentes (valor obtido pela função de equivalência da seção 4.2), 1 indica que o questionário possui uma ou mais perguntas semelhantes (valor obtido pela função de semelhança da seção 4.2) e, 0 quando o questionário não é relevante. Nota-se que 12% das consultas no modelo vetorial tem o mesmo DCG que o esperado, de acordo com o *ground truth*. Contudo, 52% das consultas realizadas no modelo vetorial apresentam uma diferença média em torno de 30% do que é esperado para o DCG, ou seja, questionários considerados fortemente relevantes que devem estar melhor ordenados estavam classificados pelo modelo vetorial em posições posteriores a questionários não relevantes e questionários moderadamente relevantes. No valor de MAP (*Mean Average Precision*), observa-se que o desempenho médio do modelo vetorial é de 56,40%. Portanto, de forma geral, constataram-se as seguintes observações nas consultas com o modelo vetorial:

- Na Consulta 1 (Figura 4), tem-se apenas 2 questionários relevantes, sendo que o primeiro questionário relevante contém perguntas com alternativas. Mas o resultado da busca retornou como sendo a posição 5, sendo que os 4 primeiros retornados não tinham relevância considerando os critérios descritos na seção 4. Isto ocorreu, provavelmente, porque o modelo vetorial não considera a relação existente entre a pergunta com a alternativa, resultando em uma precisão de 0,2. Algo similar acontece em outras 5 consultas.

- Na Consulta 2 (Figura 4), verificou-se que haviam 5 questionários relevantes de acordo com as regras definidas na seção 4, e apesar da precisão ser de 100 % para os 4 primeiros questionários ordenados, o último relevante ficou apenas na posição 88 da ordenação. A causa provável é que o vetor de palavras (qual, motivo, sua, viagem) tem o os pesos calculados pela frequência dessas palavras. Mas em se tratando de questionários, as palavras motivo e viagem deveriam ter um

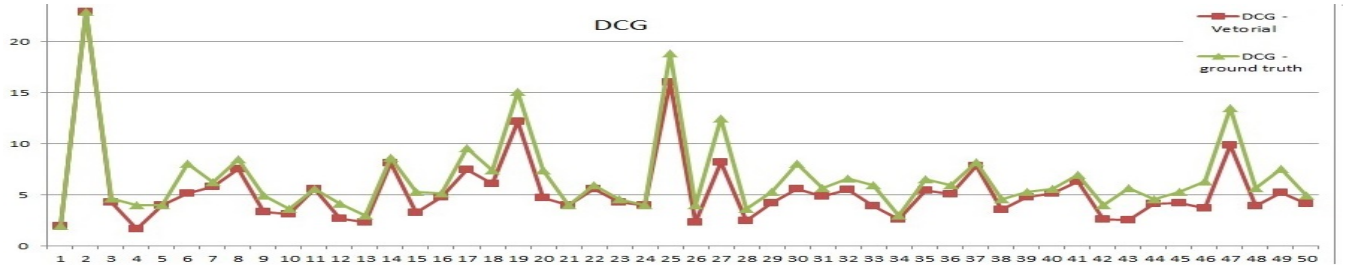

Figure 6: DCG

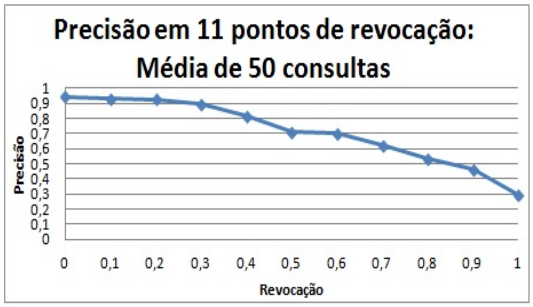

Figure 7: Precisão nos 11 pontos de revocação, considerando a média das 50 consultas

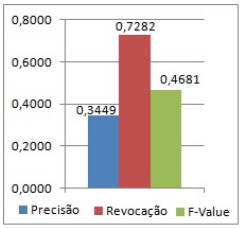

Figure 8: F-value

peso maior, devido a informação que se deseja obter com a pergunta, conforme discutido na seção 4

- Consulta 3 (Figura 4), verificou-se que haviam 4 questionários relevantes e que apesar da precisão ser de 100 % para os 3 primeiros questionários ordenados, o último relevante ficou na posição 12 da ordenação (precisão de 0,33). A causa provável é que no modelo vetorial considera-se a frequência dos termos como critério para a similaridade, e os questionários que estavam entre as posições 4 e 11 continham uma frequência maior da palavra "qual". Porém, por se tratar de questionário, a palavra "qual" é comum em perguntas, interferindo na ordenação realizada pelo modelo vetorial. Em outras 6 consultas, aconteceu algo similar.

- Consulta 4 (Figura 4), esperavam-se 9 questionários relevantes. Na consulta, após o segundo relevante, a precisão foi diminuindo drasticamente até ficar em 0,18, sendo que o ultimo relevante apareceu na posição 48. Ao analisar o parâmetro de busca, encontramos as palavras "tempo" e "posição", que conforme discutido na seção 4, não indicam qual a informação que se deseja com a pergunta (são apenas qualificadores da informação desejada), portanto, deve ter influenciado o resultado da busca, pontuando melhor questionários não relevantes do que os relevantes. Em outras 25 consultas ocorre algo similar.

O resultado da análise do experimento revelou que o modelo vetorial foi eficaz na ordenação do primeiro questionário relevante na maioria das consultas realizadas. Contudo, houve uma variação entre os resultados obtidos e os esperados na ordenação dos questionários que deveriam aparecer a partir da segunda posição em diante. Pode-se dizer que existem pontos a serem melhorados para aumentar a eficácia na recuperação de questionários.

6. TRABALHOS RELACIONADOS

O desenvolvimento de tecnologias para busca e ordenação de questionários de pesquisa ainda é um grande desafio. No entanto, é possível encontrar na literatura, alguns trabalhos relacionados a sistemas CQAs, os quais possuem características que podem ser utilizada na busca e ordenação de questionários, e portanto, podem ser analisados sob a mesma perspectiva, conforme mostrado a seguir.

Alguns trabalhos, tais como o de Chen *et al* [5], Kim *et al* [12] realizam a ordenação de perguntas de acordo com regras de definidas para determiner a relevância. Assim, com a ordenação das perguntas é possível verificar a correlação entre os documentos (que representam diferentes fóruns) e a partir do grau de correlação obtido, gerar a ordenação dos documentos relevantes. Também existe a possibilidade de ordenar as respostas de uma dada pergunta [5, 12]. Além dos trabalhos sobre sistemas CQAs, também existem trabalhos dentro da recuperação de informação, que embora sejam para recuperar documentos no geral, como por exemplo o trabalho de Pôssas *et al* [19].

Contudo, os trabalhos [16, 5, 10, 19, 21, 17, 9] que tratam a recuperação e ordenação de perguntas utilizam algoritmos voltados apenas para recuperar e ordenar perguntas do tipo aberta. Porém, como pode ser observado neste experimento, questionários de pesquisa possuem em sua maioria perguntas com alternativas (fechadas e de multipla escolha). Os trabalhos também não consideram o conjunto de perguntas nos algoritmos, ou seja, não se preocupam com a recuperação de questionários em si.

7. CONCLUSÃO

No presente artigo foi executado um experimento para verificar como o modelo vetorial se comporta na recuperação de questionários. Neste sentido foram coletados questionários na internet de modo a compor a base de dados.

Após a execução de 50 consultas, notou-se que, apesar de ser usada parte de um questionário já existente na base de dados como consulta, a precisão no primeiro ponto de revocação ficou ligeiramente abaixo do 1 (Figura 7), indicando que algumas consultas retornaram um questionário não relevante como primeiro da lista de ordenação. Nota-se também que, após o quarto ponto de revocação, há uma queda na precisão, sendo que no último ponto de revocação, a precisão chega perto de 0,2, considerando a média das 50 consultas realizadas. Por outro lado, o modelo vetorial se comporta adequadamente, considerando os primeiros questionários relevantes que devem ser recuperados, conforme pode ser observado nos primeiros 4 pontos de revocação.

Como trabalho futuro, pretende-se formalizar as funções de equivalência, semelhança e ausência, realizar uma alteração no cálculo dos pesos de forma a melhorar o resultado. Pretende-se melhorar o cálculo da similaridade considerando a correlação entre pergunta e alternativas. Por fim, realizar experimentos para averiguar se haverá melhoria na ordenação ao realizar um processo de tratamento nos termos que indicam que a sentença é uma pergunta (exemplo qual, porque, onde, quando.). Vale ressaltar que é interessante realizar experimentos com outros modelos e analisar qual seria o modelo mais indicado para ser utilizados com recuperação e ordenação de questionários.

8. ACKNOWLEDGMENTS

Nossos agradecimentos ao aluno Gilnei da UFSC que desenvolveu o *crawler* para buscar questionários em seu trabalho de conclusão de curso.

9. REFERENCES

[1] Apache lucene. https://lucene.apache.org/. (Accessed: 01-2017), 2017.

[2] M. M. d. Andrade et al. Introdução à metodologia do trabalho científico, 1999.

[3] Y. BAEZA and B. Ribeiro-Neto. Modern information retrieval-the concepts and technology behind search, 2011.

[4] N. G. Barnes. 2013 fortune 500-umass dartmouth. http://www.umassd.edu/cmr/socialmediaresearch/2013fortune500/. (Accessed: 09-03-2017).

[5] R.-C. Chen, D. Spina, W. B. Croft, M. Sanderson, and F. Scholer. Harnessing semantics for answer sentence retrieval. In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 21–27. ACM, 2015.

[6] L. P. Claudino, M. B. Nunes, and F. d. Silva. Finanças pessoais: um estudo de caso com servidores públicos. *Anais do SEMEAD-Seminários em Administração, São Paulo, SP, Brasil*, 12, 2009.

[7] M. T. V. Coelho. *Comunicação terapêutica em Enfermagem: utilização pelos enfermeiros*. PhD thesis, Instituto de Ciências Biomédicas Abel Salazar, 2015.

[8] M. C. de Souza Minayo. *Pesquisa social: teoria, método e criatividade*. Editora Vozes Limitada, 2011.

[9] A. Grappy, B. Grau, M.-H. Falco, A.-L. Ligozat, I. Robba, and A. Vilnat. Selecting answers to questions from web documents by a robust validation process. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 1, pages 55–62. IEEE, 2011.

[10] E. H. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. Question answering in webclopedia. In *TREC*, volume 52, pages 53–56, 2000.

[11] S.-m. Kim and Y.-g. Ha. Automated discovery of small business domain knowledge using web crawling and data mining. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 481–484. IEEE, 2016.

[12] S. N. Kim, L. Cavedon, and T. Baldwin. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics, 2010.

[13] N. N. Knupfer and H. McLellan. 41. descriptive research methodologies. 1996.

[14] C. R. Kothari. *Research methodology: Methods and techniques*. New Age International, 2004.

[15] E. M. Lakatos and M. d. A. Marconi. Fundamentos da metodologia científica. In *Fundamentos da metodologia científica*. Altas, 2010.

[16] S. L. Lo, R. Chiong, and D. Cornforth. Ranking of high-value social audiences on twitter. *Decision Support Systems*, 85:34–48, 2016.

[17] P. Molino and L. M. Aiello. Distributed representations for semantic matching in non-factoid question answering. In *SMIR@ SIGIR*, pages 38–45, 2014.

[18] C. F. Picard. *Graphs and questionnaires*, volume 32. Elsevier, 1980.

[19] B. Pôssas, N. Ziviani, W. Meira Jr, and B. Ribeiro-Neto. Set-based vector model: An efficient approach for correlation-based ranking. *ACM Transactions on Information Systems (TOIS)*, 23(4):397–429, 2005.

[20] P. B. Sheatsley, P. H. Rossi, J. D. Wright, and A. B. Anderson. Questionnaire construction and item writing. *Handbook of survey research*, pages 195–230, 1983.

[21] W. Song, M. Feng, N. Gu, and L. Wenyin. Question similarity calculation for faq answering. In *Semantics, Knowledge and Grid, Third International Conference on*, pages 298–301. IEEE, 2007.

[22] I. Srba and M. Bielikova. A comprehensive survey and classification of approaches for community question answering. *ACM Trans. Web*, 10(3):18:1–18:63, Aug. 2016.

[23] S. Vieira. *Como elaborar questionários*. Atlas, 2009.

[24] L. M. Villar, A. J. d. Almeida, M. C. A. d. Lima, J. L. V. d. Almeida, L. F. B. d. Souza, and V. S. d. Paula. A percepção ambiental entre os habitantes da região noroeste do estado do rio de janeiro. *E. Anna Nery Revista Enfermagem*, 12(2):285–290, 2008.

[25] R. S. Waslawick. *Metodologia de pesquisa para ciência da computação*. Elsevier, Rio de Janeiro, 2014.

[26] D. Yang, M. Piergallini, I. Howley, and C. Rose. Forum thread recommendation for massive open online courses. In *Educational Data Mining 2014*, 2014.