QASF 2.0 - SISTEMA DE PERGUNTAS E RESPOSTAS EM DOENÇAS CRÔNICAS

QASF 2.0 - QUESTION-ANSWERING SYSTEM ON CHRONIC DISEASES

QASF 2.0 - SISTEMA DE PREGUNTA Y RESPUESTA EN ENFERMEDADES CRÓNICAS

Gabriel Miranda Rubio[1], Gilvan Machado Morais[1], Ester E. Domingues dos Santos[1], Renato de Freitas Bulcão-Neto[2] e Alessandra Alaniz Macedo[1]

[1]DCM-FFCLRP/Universidade de São Paulo (USP), Ribeirão Preto-SP, Brasil
[2] Universidade Federal de Goiás (UFG), Goiânia-GO, Brasil

Resumo: As doenças crônicas são um grave problema mundial de saúde pública. Elas são responsáveis por morbidez, mortalidade, incapacidade e baixa qualidade de vida. No Brasil, 72% das mortes são causadas por doenças crônicas, e no mundo esse valor é de 60%. Sabe-se que a atual produção cientifica mundial é vasta e as informações contidas nos artigos científicos podem ser úteis para o profissional da saúde. Porém, ler artigos científicos é uma tarefa que exige tempo e disposição. O objetivo deste artigo é apresentar a segunda versão do sistema de perguntas e respostas QASF para auxiliar os profissionais da saúde e outros usuários na busca rápida por informações, especificamente, em doenças crônicas. O processo de desenvolvimento do sistema baseou-se na construção de módulos de processamento de pergunta, de documentos e de resposta. Com o QASF, espera-se facilitar o acesso a informações.

Palavras-chave: Doença Crônica, Mineração de Texto, Computação em Informática Médica.

***Abstract**: Chronic diseases is a global public health problem. They are responsible for morbidity, mortality, disability and poor quality of life. Chronic diseases cause 60% of deaths worldwide and 72% of deaths in Brazil. It is known that the current world scientific production is vast and the information contained in scientific papers can be useful for health professionals. However, the reading of scientific articles is a demanding and time-consuming task. This paper presents a question answering system to help health professionals quickly search for information about chronic diseases. The core of the question answering system includes the development of modules for processing questions, documents and answers. It is expected to support users that demand for quick information about chronic diseases. Precision rates between 50% and 60% were obtained.*

***Keywords**: Chronic Disease, Textual Mining, Medical Informatics Computing.*

Introdução

As doenças crônicas são um sério problema mundial de saúde pública. Elas são responsáveis por morbidez, mortalidade, incapacidade e baixa qualidade de vida[1], sendo responsável por 72% das mortes no Brasil[2] e por 60% de mortes no mundo[3]. Sabe-se que a produção cientifica atualmente é imensa e as informações contidas nos artigos científicos podem ser úteis para o profissional da saúde. Porém, ler artigos científicos é uma tarefa que exige tempo e disposição.

Os sistemas de perguntas e respostas (QA - *Question Answering*) têm como objetivo fornecer informações diretas e precisas sobre uma pergunta proposta pelo usuário do sistema. Para realizar esta tarefa, sistemas de QA utilizam técnicas computacionais, principalmente, das áreas de Extração de Informação, Processamento de Linguagem Natural e Recuperação de Informação[4].

Nos sistemas tradicionais de busca (do inglês, *search engines*), um usuário deve compor uma consulta com termos selecionados e conectores lógicos para receber um conjunto de links para documentos completos como resposta.

Quando o usuário faz buscas por informações de interesse nos sistemas de QA, o primeiro passo é o usuário elaborar uma questão em linguagem natural que traduza a sua necessidade de busca. Dependendo da complexidade dos sistemas de QA, os sistemas podem responder a vários formatos de perguntas. O formato de pergunta mais comum e simples é o de perguntas que se baseiam em fatos (factual), por exemplo, "Quando foi produzido o primeiro antibiótico?". Com o avanço da área, outros tipos de perguntas mais complexas puderam ser incluídas como, por exemplo, as questões com múltiplas respostas ("Quais são os sintomas da tuberculose?"), as questões de definição ("O que é câncer?") e as questões do tipo "Por que" ("Por que a hipertensão pode causar problemas?")[4].

O objetivo deste artigo é apresentar o desenvolvimento do sistema de perguntas e respostas (QASF - *Question Answering System*) - *Version* 2.0 para suportar perguntas específicas e diretas de profissionais da saúde, quando analisando pacientes com suspeita de fatores de risco para o desenvolvimento de doenças crônicas, ou mesmo fazendo consultas científicas. O sistema QASF segue a arquitetura da Figura 1 proposta em[6]. Os sistemas apoiados pelo QASF, como os tradicionais sistemas de QA[5], apresentam uma arquitetura básica composta por três módulos principais: (i) Processamento da Pergunta (PP), (ii) Processamento do Documento (PD) e (iii) Processamento da Resposta (PR). O módulo PP é composto pelas tarefas de "Elaboração da Consulta" que utiliza as técnicas de processamento de linguagem natural para a eliminação de *stopwords*, ou seja, eliminação de palavras comuns em um determinado documento, e "Análise da Questão" que ajuda na ordenação e na seleção das respostas no módulo de PR. O módulo de PD recebe a pergunta processada pelo módulo PP e recupera uma possível resposta em resumos de artigos científicos da biblioteca digital online *PubMed*. Finalmente, o módulo de PR realiza a "Ordenação e Seleção das Respostas" selecionando as frases do artigo relacionadas à pergunta por meio de cálculos de similaridade (cosseno), e faz a "Validação das Respostas".

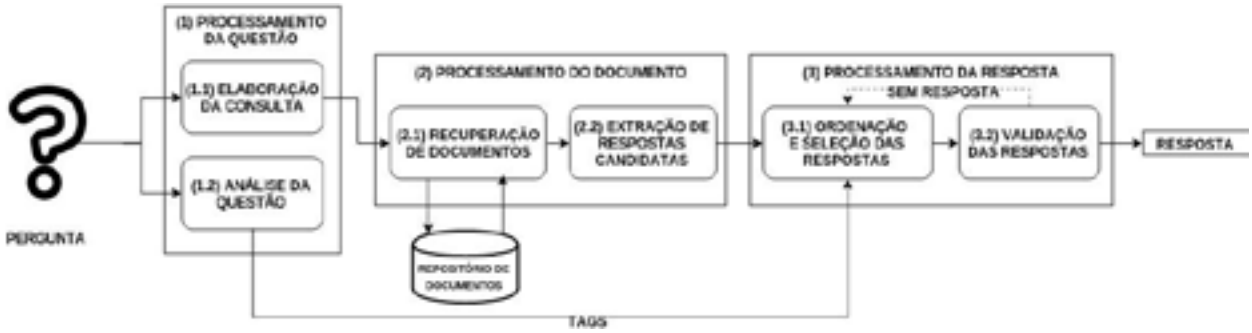

Figura 1 - Arquitetura do sistema de QA[6]

O restante do artigo está organizado da seguinte maneira: a Seção 2 apresenta o sistema QASF 2.0; a Seção 3 detalha materiais e métodos usados para o desenvolvimento de cada módulo de processamento do QASF 2.0 e finaliza com o algoritmo do sistema; a Seção 4 apresenta os resultados e a discussão; e finalmente, a Seção 5 conclui com o artigo e apresenta trabalhos futuros.

O Sistema QASF 2.0

A versão 2.0 do sistema QASF foi desenvolvida para ser um sistema *Web* rápido e simples de usar. As interfaces do sistema que compõem um *site* foram desenvolvidas com o framework Django versão 1.9, usando a linguagem Python com HTML5. O *site* é expansível e modularizado, uma vez que cada módulo desenvolvido no projeto pode ser introduzido de maneira fácil no sistema ou em outro sistema, sem perder desempenho no processamento e com layout simples.

A Figura 2 apresenta a tela inicial do site. Essa interface é simples para ser de fácil manuseio por usuários inexperientes. O único campo textual é central e suporta a inserção da pergunta de usuários na forma de consulta em linguagem natural. Essa consulta é enviada aos módulos de processamento apresentados na Seção 3. Abaixo do campo de consulta, existem opções de tratamento da consulta que têm opções *default* pré-selecionadas ou podem ser modificadas pelos usuários. Após o envio da consulta, a página é redirecionada para a tela de apresentação das respostas (ver Figura 3).

Figura 2: Tela inicial, local onde a pergunta é inserida.

Figura 3: Tela de apresentação das respostas com trechos de artigos relacionados com a resposta

A Figura 3 ilustra a tela de apresentação das respostas e é formada pelas possíveis respostas retornadas dos módulos de processamento. No exemplo proposto, a pergunta "*What are the adult health problems of a kid with malnutrition indication, short stature, neuropsychomotor development appropriate, updated vaccination, low fiber diet, an appropriate childhood routine, some family members who have chronic illness, violent emotional environmental, lives in a flood area, with a lack of family entertainment, with risk of behavioral disorders, with physical exercises practice, with asthma, with pos-term of 42 completed weeks and coronary artery disease on the family history*?" foi inserida, e o sistema redirecionou o usuário para a página da Figura 3 contendo a apresentação das respostas.

Materiais e métodos

No sistema QASF 2.0, a criação de perguntas pelo usuário ocorre usando linguagem natural. Em seguida, o módulo de Processamento de Pergunta (PP) realiza a análise e a classificação da pergunta e a formulação de uma consulta ao banco de dados criado com os resumos dos artigos do sítio do *PubMed* sobre doenças crônicas. A consulta gerada pelo módulo PP é utilizada para recuperar os resumos relevantes. Dos resumos mais relevantes recuperados, trechos candidatos para atender à resposta são extraídos no módulo Processamento de Documento (PD) e utilizados como entrada pelo módulo de Processamento de Resposta (PR). Para a seleção de respostas, os trechos candidatos a resposta são comparados com o tipo de pergunta feita e ordenados de acordo com a similaridade com a pergunta. O módulo PR utiliza esses trechos para compor um documento de resposta. O profissional da saúde poderá realizar perguntas relacionadas a doenças crônicas em pesquisas científicas ou a doenças crônicas relacionadas a fatores de risco do paciente. Os módulos PP, PD e PR são detalhados a seguir.

Análise de Processamento de Perguntas (PP) - Para análise do processamento de perguntas em linguagem natural são realizados os seguintes processamentos:

- Eliminação de *Stopwords.* As *stopwords* são palavras comuns (ex. preposições, artigos e pronomes) do vocabulário que, dependendo do tipo de busca, são descartadas, ou seja, são retiradas antes ou durante o processamento de texto em linguagem natural. Existem prós e contras desse processamento. A retirada de *stopwords* simplifica a análise e possui o intuito de melhorar o desempenho das buscas. Nos casos dos sistemas de QA, a retirada total dessas palavras impediria a distinção de nomes que usam *stopwords* em forma de entidade representativa. Esse processamento pode prejudicar a busca por sentenças e um conjunto de textos que possua um tema comum. Alguns documentos possuem palavras que podem ser vistas em uma distribuição uniforme, mesmo que sejam *stopwords* que nesses casos possuem um significado importante. No processamento da pergunta foi utilizado o conjunto de *stopwords* definido no *NLTK*[11] (*Natural Language Toolkit*) (Figura 4 – Linha 09). Como resultado, a maioria das palavras do tipo substantivos, verbos, adjetivos ficam isoladas (*bag of words*) para formar um vetor de consulta. Por exemplo:
    - PERGUNTA DE ENTRADA: *"What complications increase of the Elective caesarean section (CS) before 39 weeks?"*
    - VETOR-PERGUNTA DE SAÍDA: *"['complications', 'increase', 'elective', 'caesarean', 'section', '(cs)', '39', 'weeks']"*
- Etiquetamento lexical. O *pos-tag*, ou *part-of-speech tagger*, é um processo de fixar para cada palavra uma etiqueta (do inglês, *tag*), com o intuito de dividir o discurso em partes, as quais são classes de palavras ou categorias lexicais. O processamento em *NLTK* usa convenções diferentes para etiquetar palavras. Essas etiquetas são conhecidas como *tagset (ver Apêndice A).* O *pos−tag* foi utilizado para a classificação lexical das palavras dentro de uma frase. Esse processamento identifica as classes de palavras mais relevantes para a consulta do usuário em

linguagem natural, reconhecendo as palavras-chave de acordo com sua importância lexical. O etiquetamento permitiu a escolha das etiquetas que são mais importantes para cada processamento (Figura 4 - Linhas 11, 14 e 15). Porém, a liberdade de escolha das etiquetas pode fazer com que etiquetas erradas sejam descartadas e informações importantes sejam perdidas. Usando a pergunta *"What complication sincrease of the Elective caesarean section (CS) before 39 weeks?"*, após o processo de relacionar palavras com as *tagset*, a pergunta seria transformada em:

- "[('*What*', 'WP'), ('*complications*', 'NNS'), ('*increase*', 'VBP'), ('*of*', 'IN'), ('*the*', 'DT'), ('*Elective*', 'NNP'), '*caesarean*', 'NN'), ('*section*', 'NN'), ('(CS)', 'NN'), ('*before*', 'IN'), ('39', 'CD'), ('*weeks*?', 'JJ')]"

- Busca de definições no *Wordnet*. *O WordNet*[9] é um banco de dados lexical (com substantivos, verbos, adjetivos e advérbios) disponível para *download*. Ele é similar a uma enciclopédia, que relaciona uma palavra ao seu significado ou sentido específico. Porém, o *WordNet* possui apenas definições especificas. No QASF 2.0, foram utilizadas duas abordagens (Figura 4 - Linhas 05-07): (1) antes da pergunta ser realizada pelo usuário, ela é processada por um dos dois módulos, a critério do usuário, e (2) a pergunta é transformada a partir de um dicionário na tentativa de encontrar a resposta, sem a necessidade de processamento mais elaborado. A segunda abordagem utiliza uma biblioteca do *WordNet* com uma função de pesquisa por hipertermos (do inglês, *hyperterms*). Esses termos são inseridos na busca, para uma procura mais abrangente para o caso da palavra que está sendo pesquisada não ser encontrada nos artigos pesquisados. Por exemplo:
  - Pergunta inserida pelo usuário: *"Whatis diabetes mellitus?"*.
  - Durante o processamento (2) é feita uma busca no *WordNet* para as palavras: "*diabetes mellitus"*.
  - Resposta retornada pelo dicionário *WordNet*: *"diabetes caused by a relative or absolute deficiency of insulin and characterized by polyuria".*

A priori, a pergunta é processada pelo dicionário. Caso nenhuma resposta seja encontrada, o processamento mais elaborado é acionado, e utilizando hipertermos, a *query* é reformulada, formando assim uma matriz contendo também termos mais abrangentes como apresentado:

- Pergunta inserida pelo usuário: *"What contributes to the development of microvascular complications of diabetes mellitus?"*.
- Nenhuma resposta foi encontrada passando pelo processamento (2).
- *Query* é reformulado com os hipertermos gerando a seguinte matriz:
  - [[,*What*'], [,*change*', ,*contribute*'], [,*to*'], [,*the*'], [,*improvement*', ,*development*'], [,*of*'], [,*microvascular*'], [,*hindrance*', ,*complications*'], [,*of*'], [,*polygenicdisorder*', ,*diabetes*'], [,*mellitus*']]".
- Em seguida, a query é submetida o processamento, selecionado pelo usuário.

Análise do Processamento de Documento (PD) - O módulo de Processamento de Documentos (PD) recebe a pergunta processada no módulo PP para recuperar uma possível resposta em artigos científicos de bibliotecas digitais online como o *PubMed*. Atualmente, o sistema QASF utiliza artigos científicos do *PubMed* filtrados por uma ontologia de Doenças Crônicas para formar uma coleção curada e específica de artigos científicos do domínio. A versão 2.0 do QASF está consultando apenas os resumos dos artigos. A coleção consultada possui 1097 artigos com 1097 resumos com a média de 345 palavras por resumo de artigo. Os artigos possuem tags que marcam o título do artigo (TI), o código do artigo (PMID), o abstract (AB) e outras tags, que a priori, não foram consideradas para o processamento. As tags TI, PMID, AB foram utilizadas pelo sistema.

O processamento do documento (resumo do artigo) ocorreu da seguinte maneira: cada artigo foi selecionado e inserido em uma lista. Um pré-processamento foi realizado para conversão de todas as

letras do documento para em minúsculo. Em seguida, são realizados cálculos de cosseno pelo módulo de Processamento de Resposta (apresentado na Seção 3.3) para atribuir similaridade entre a informação resultante do módulo PP para representação da pergunta e o documento processado pelo módulo PD.

A escolha do tipo de cosseno no momento da inserção da pergunta (na Figura 2) pelo usuário interfere neste módulo de PD, uma vez que a saída desse módulo depende do tipo de cosseno escolhido para ser usado. No caso do "Cosseno com *Pos-tag*", a saída do processamento de documentos fica na forma de lista, sendo que cada palavra possui uma etiqueta lexical. Já no caso do cosseno chamado de "Cosseno Algébrico", [12], os documentos ficam também em forma de lista, mas sem as etiquetas lexicais.

Análise de Processamento de Resposta (PR) - O processamento de resposta relaciona uma pergunta feita em linguagem natural com um trecho de um documento, que possui uma certa similaridade, e possivelmente seja a resposta para a pergunta. Esse processamento utiliza a medida de similaridade do cosseno entre dois vetores para indicar a similaridade da pergunta do usuário com trechos dos artigos científicos do *PubMed*. Os cálculos de similaridade são do QASF 2.0:

- Cossenos com *Pos-tag* (Figura 4 - Linhas 13-20). Cada palavra do documento recebe uma etiqueta, similar às usadas no Etiquetamento lexical, deixando prontas as lista de documentos para o Processamento de Resposta, que a usa. Quando uma palavra da pergunta recebe uma etiqueta, é realizada uma intersecção de etiquetas considerando que palavras iguais, na pergunta e na resposta, tenham a mesma etiqueta. Uma lista com as etiquetas da pergunta e de todas as possíveis respostas é gerada, e em seguida é realizada uma contagem para o cálculo do cosseno entre os vetores gerados da pergunta e da resposta. Por exemplo, para a pergunta "*What complications increase of the Elective caesarean section (CS) before 39 weeks?*", o sistema gera a seguinte lista de etiquetas "WP NNS VBP IN DT NNP NN NNNN IN CD JJ". E a resposta seria "CD : JJ NN NN TO VB JJ NN NN CC JJ MD VB IN NN NNNN NNP JJ NNS IN CD NNS NNS VBP JJ NN NN CC JJ NN NN NNP JJ NN NN IN NNS NN IN NNS NNS IN VBN NN IN NN NNNN CC RB NN CD NNP JJ NNS MD VB JJ : IN NNS : IN NN NNP NNP NN IN JJ NNS MD VB IN NN JJ JJ NNP". Quando cada etiqueta é contabilizada gera-se:
  - Pergunta: *„Counter({,NN': 3, ,IN': 2, ,WP': 1, ,CD': 1, ,VBP': 1, ,JJ': 1, ,DT': 1, ,NNS': 1, ,NNP': 1})"*
  - Resposta: *„Counter({,NN': 22, ,JJ': 12, ,IN': 10, ,NNS': 9, ,NNP': 6, ,VB': 4, ,MD': 3, ,CC': 3, ,CD': 3, ,RB': 1, ,VBN': 1, ,TO': 1, ,VBP': 1})"*.
- Cossenos Algébrico (Figura 4 - Linhas 21-27). O cosseno Algébrico não requer processamento de etiquetamento, visto que o cosseno é calculado usando a contagem de palavras de cada documento da lista de documentos. O cosseno Algébrico é similar ao anterior, porém o cosseno é calculado usando a contagem de palavras feita na pergunta e na resposta. Para isso, um vetor intersecção é calculado. Também é necessário o cálculo do produto escalar e a identificação das dimensões em comum dos vetores. Desse modo, não é necessário a normalização do vetor. Por exemplo, tem-se:
  - Pergunta: *"Counter({'What': 1, 'weeks': 1, 'of': 1, 'caesarean': 1, '39': 1, 'increase': 1, 'complications': 1, 'Elective': 1, 'CS': 1, 'the': 1, 'section': 1, 'before': 1})"*
  - Resposta: *„Counter({,cs': 4, ,and': 3, ,for': 3, ,in': 2, ,maternal': 2, ,complications': 2, ,among': 1, ,food': 1, ,sepsis': 1, ,into': 1, ,stillbirth': 1, ,primary': 1, ,born': 1, ,epigenetic': 1, ,accounts': 1, ,should': 1, ,occur': 1, ,unit': 1, ,before': 1, ,1': 1, ,different': 1, ,asthma': 1, ,allergy': 1, ,decision': 1, ,to': 1, ,of': 1, ,take': 1, ,weeks': 1, ,diabetes': 1, ,type': 1, ,mortality': 1, ,elective': 1, ,be': 1, ,repeat': 1, ,increased': 1, ,uterine': 1, ,may': 1, ,caesarean': 1, ,incidence': 1, ,possible': 1, ,colonization': 1, ,admissions': 1, ,possibly': 1,*

‚increases‘: 1, ‚neonates‘: 1, ‚others‘: 1, ‚might‘: 1, ‚consideration‘: 1, ‚previa‘: 1, ‚rupture‘: 1, ‚by‘: 1, ‚21950660ab‘: 1, ‚care‘: 1, ‚placenta‘: 1, ‚neonatal‘: 1, ‚39‘: 1, ‚respiratory‘: 1, ‚hypoglycaemia‘: 1, 1, ‚morbidity‘: 1, ‚section‘: 1, ‚responsible‘: 1, ‚consequent‘: 1, ‚intensive‘: ‚increta‘: 1, ‚gut‘: 1, ‚malignancies‘: 1, ‚conclusion‘: 1, ‚changes‘: 1, ‚childhood‘: 1, ‚foetal‘:1})“

O valor do cosseno para a abordagem que usa Cosseno Algébrico do exemplo é 0.27. O exemplo do Cosseno com *Pos-Tag* é 0.64.

Algoritmo do QASF 2.0 - A Figura 4 apresenta um algoritmo que ilustra os principais processamentos do QASF 2.0. As linhas de 5 a 12 apresentam o Processamento de Pergunta. As linhas de 13 a 26 detalham o Processamento de Resposta. O detalhamento do Processamento de Documento da Seção 3.2 não foi incluso no algoritmo por motivos de simplificação.

```
01. Inicio
02.         Input: pergunta // do usuario em linguagem natural
03.         PP(pergunta,opp) // pergunta transformada em termos opp
04.         Escolhe tipo_cosseno(tc)
05.         Se pergunta ∈ WordNet definitions
06.         |     Retornar resposta
07.         Senão
08.         |     Se opp é "Stopword"
09.         |     |     pergunta <= pergunta(not stopword)
10.         |     Senão
11.         |     |     pergunta <= tag_opp(pergunta) //termos com tags especificas
12.         |     Fim_se
13.         |     Se tc é "cosseno_pos_tag"
14.         |     |     lista_pergunta <= tag(pergunta)
15.         |     |     lista_documentos <= tag(documentos)
16.         |     |     vetor_documento <= contagem de lista_documentos
17.         |     |     vetor_pergunta <= contagem de lista_pergunta
18.         |     |     Para cada vetor em vetor_documento
19.         |     |     |     lista_respostas <= cosseno(vetor_pergunta, vetor) //cálculo entre
            |     |     |                                   //vetor_documento e vetor resposta
20.         |     |     Fim_para
21.         |     Senão
22.         |     |     vetor_documento <= contagem de palavras dos documentos
23.         |     |     vetor_pergunta <= contagem de palavras da pergunta
24.         |     |     Para cada vetor em vetor_documento
25.         |     |     |     lista_respostas <= cosseno(vetor_pergunta, vetor) //cálculo entre
            |     |     |                                   //vetor_documento e vetor resposta
26.         |     |     Fim_para
27.         |     Fim_se
28.         Fim_se
29.         Retorna  max(lista_respostas) //documento com maior valor de cosseno
30. Fim
```

Figura 4: Algoritmo do sistema de QA

Resultados e Discussão

Para experimentação do QASF 2.0, foram realizados experimentos com resultados apresentados

nesta seção. Devido à incompatibilidade do exíguo tempo de projeto e a realização de trâmites em Comissão de Ética para permissão do uso oficial do sistema por profissionais da sáude, inicialmente, foram realizadas simulações de uso do sistema. Com esse propósito, foram criadas perguntas com base em registros médicos de pacientes cujos prontuários haviam sido relacionados com artigos do *PubMed* e apresentados em trabalhos prévios do grupo8. Esses relacionamentos foram classificados em fortemente relevantes, relevantes, pouco relevantes e irrelevantes. Os cenários de experimentação, a partir das validações de processamento, são:

- Perguntas passadas sem tratamento pelo Cosseno com *Pos-tag*;
- Perguntas passadas sem tratamento pelo Cosseno Algébrico;
- Perguntas retirando as *Stopwords* e usando Cosseno com *Pos-tag*;
- Perguntas retirando as *Stopwords* e usando Cosseno Algébrico;
- Perguntas passando por tratamento de *Pos-tag* pelo Cosseno com *Pos-tag*;
- Perguntas passando por tratamento de *Pos-tag* pelo Algébrico.

Para cada cenário, foram calculadas três medidas de precisão para cada grupo de artigos científicos selecionados: (i) os 3 maiores cossenos entre as perguntas e os três artigos científicos retornados, (ii) os cinco primeiros com maior cosseno e (iii) os dez primeiros com maior cosseno. Logo após, foram verificados se esses artigos retornados se encontravam entre os que eram classificados fortemente relevantes ou relevantes. Se o artigo estava em uma dessas duas classificações, ele foi considerado como um acerto. O Quadro 1 apresenta os resultados obtidos em termos de contagem de acertos para cada cenário. As colunas apresentam os tipos de processamento (os cenários) e as linhas apresentam a quantidade de artigos científicos selecionados (3, 5 e 10 primeiros com maior valor de cosseno). Finalmente, as células do quadro apresentam as precisões.

Tabela 1: Medidas de precisão relacionadas a cada cosseno

Qtde de artigos retornados	(1) Cos_pos_tag	(2) Cos_Alg	(3) Cos_ pos_tag_ stopWord	(4) Cos_Alg_ StopWord	(5) Cos_pos_ tag_pos_tag	(6) Cos_Alg_ pos_tag
3 primeiras	0,2	0,1	0,1	0,6	0	0,6
5 primeiros	0,2	0,2	0,1	0,5	0	0,5
10 primeiros	0,1	0,2	0,1	0,5	0	0,5

Na Tabela 1, pode ser visto que os melhores valores de precisão ocorreram nos cenários (4) e (6) de "Perguntas retirando as *Stopwords*e usando Cosseno Algébrico" e de "Perguntas passando por tratamento de *Pos-tag* pelo Cosseno Algébrico" respectivamente. Portanto, o uso do cosseno algébrico alcançou uma melhor performance de precisão em relação aos outros. Foi possível também verificar que a frequência de artigos fortemente relevantes e relevantes aparece na maioria das vezes entre os cinco primeiros artigos retornados. A revocação dos resultados também foi calculada. Os resultados de revocação (do inglês, *recall*) e de precisão (do inglês, *precision*), considerando, as vinte e cinco primeiras respostas retornadas, são apresentados no Gráfico 1. A precisão foi calculada considerando o cenário que retornava mais documentos de forte relevância em relação a todos os documentos retornados. A revocação foi calculada considerando os documentos de forte relevância retornados com relação à quantidade de documentos de forte relevância ligados à pergunta inserida. Foi possível notar que os processamentos 4 e 6 possuem os maiores valores de precisão e revocação. Esses valores podem possivelmente ser explicados pelo fato do cálculo do Cosseno Algébrico ser mais eficiente para o sistema e a coleção manipulada. O pior caso foi o do processamento do cenário 5, uma possível razão para essa taxa pode ser que as palavras retiradas com o processamento de *Pos-tag* eram necessárias para o cálculo de cosseno *Pos-tag*. Em relação ao baixo nível geral da revocação, pode-se ser atribuído

ao alto número de documentos relevantes, uma vez que na maioria das vezes são medidas inversamente proporcionais. A baixa revocação não é preocupante para sistemas como o QASF, uma vez que se espera poucos retornos de documentos, porém precisos e com a resposta exata em destaque.

Conclusão

Nas áreas de medicina e de saúde, investigar informações contidas em prontuários médicos de pacientes junto a pesquisas em informações científicas pode contribuir para o tratamento e para a prevenção de doenças. No ambiente médico, buscar informações relevantes em artigos científicos de cada área é uma tarefa que exige tempo e experiência dos profissionais, principalmente na formulação de consultas e na navegação por diferentes bibliotecas digitais. Além disso, ler artigos científicos em busca de informações relevantes é uma tarefa dispendiosa de tempo, além das tarefas profissionais diárias. Portanto, o sistema QASF pode auxiliar os profissionais da área da saúde na busca rápida por informações precisas, por exemplo, no domínio de doenças crônicas. O processamento do sistema foi validado com diferentes abordagens.

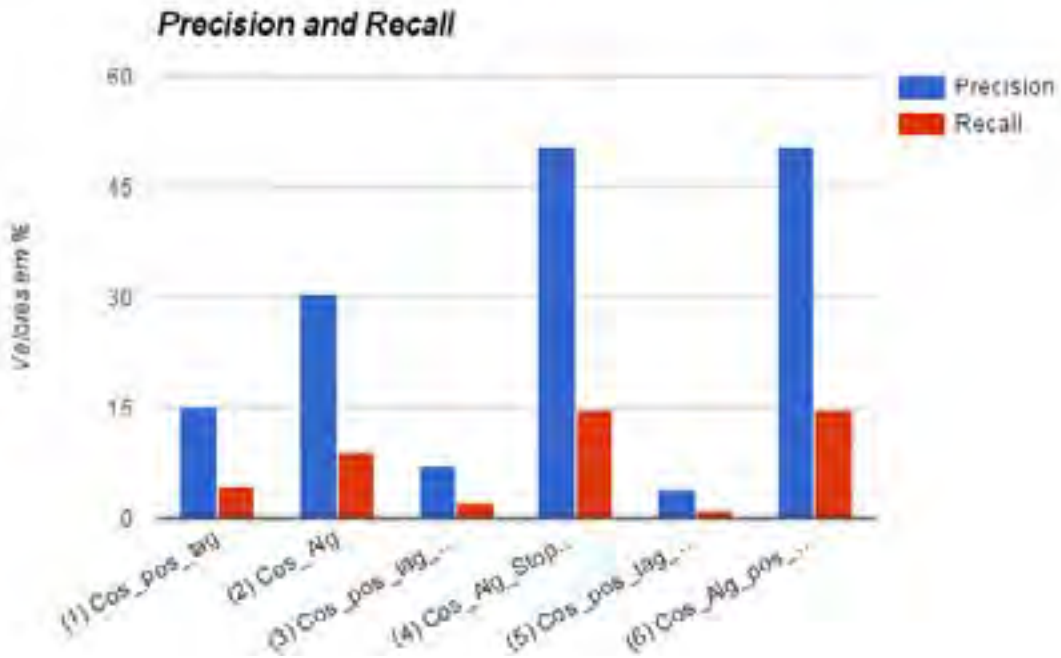

Gráfico 1: Precisão e revocação calculadas com as primeiras 25 respostas.

Com o trabalho em andamento, os autores estão realizando novas avaliações considerando outras medidas como, por exemplo, a medida F. Além disso, outros planos futuros são: (i) finalizar a inclusão dos *frameworks* SisViDas e FREDs para finalização do QASF 3.0, uma vez que esses *frameworks* processam a similaridade entre consultas enviadas, por exemplo pelo móulo PP do QASF, com documentos de domínios específicos; (ii) avaliar o QASF em outros domínios de informação como, por exemplo, o câncer de tireoide suportado pelo *framework* FREDS; (iii) transformar o QASF em uma aplicação que composta por interfaces mobile para facilitar a interação; (iv) avaliar e comparar os resultados entre os dicionários do *Wordnet* e outros dicionários da área médica; (v) incluir o uso de UMLS (*Unified Medical Language System*) e de ontologias do domínio como, por exemplo, a CDO (*chronic disease ontology*) no processamento dos textos; (vi) acrescentar uma etapa de pré-processamento da questão na etapa de "Análise da Questão" por técnicas de aprendizado de máquina contendo

abordagens como, por exemplo, a NER (*Named-Entity Recognition*); e (vii) realizar um estudo comparativo entre estas abordagens para descobrir quais técnicas agrupadas aprimorarão o desempenho.

Agradecimentos

Os autores agradecem à Pró-Reitoria da Universidade de São Paulo pelas bolsas dos alunos no projeto unificado e também a aluna de mestrado Luciana Farina Almansa pelo apoio no desenvolvimento das ideias e da primeira versão do sistema.

Referências

[1] Holt JB, Huston SL, Heidari K, Schwartz R, Gollmar CW, Tran A, Bryan L, Liu Y, Croft JB. Indicators for chronic disease surveillance - United States, 2013. MMWR Recomm Rep. 2015 Jan ;64(RR-01):1-246.

[2] Organização Mundial da Saúde. O impacto das doenças crônicas no Brasil [Internet]. [atualizado em 2015 Dec 21; citado em 2016 Mai 8].Disponível em: www.who.int/chp/chronic_disease_report/media/brazil.pdf

[3] Organização Mundial da Saúde. Chronicdiseasesandhealthpromotion [Internet].[citado em 2016 Mai 8].Disponível em: www.who.int/chp/en/

[4] Allam AMN, Haggag MH. The question answering systems: A survey. IJRRIS. 2012 Sep. 2(3): 211–20.

[5] Hirschman L, Gaizauskas R. Natural language question answering: the view from here. Nat Lang Eng. 2001 Dec. 7(4): 275–300.

[6] Almansa LF, Macedo AA. Sistema de Informação para Perguntas e Respostas em Doenças Crônicas. Em 16o Workshop de Informática Médica. Congresso da Sociedade Brasileira de Computação. 2016 Jul. Porto Alegre, Brasil. 10p.

[7] Macedo AA,Pessotti HC,Almansa LF, Felipe JC, Kimura ET. Morphometric information to reduce the semantic gap in the characterization of microscopic images of thyroid nodules.Comput Methods Programs Biomed. 2016 Jul ;130:162-74.

[8] Pollettini JT, Baranauskas JA, Ruiz ES, Pimentel MG, Macedo AA. Surveillance for the prevention of chronic diseases through information association. BMC Medical Genomics. 2014 Jan. 7(1):7.

[9]Miller GA. Wordnet - A Lexical database of English [Internet]. Princeton: Princeton University Department of Psychology, Department of Computer Science. [atualizado em 2015 may; citado em 2016 jun]. Disponível em : https://wordnet.princeton.edu/.

[10]Bird S, Klein E, Loper E. Natural Language Processing with Python [Internet].Califórnia:O'Reilly Media. 2009 June [citado em mai 2016]Disponível em : http://www.nltk.org/book_1ed/ch05.html.

[11]NLTK - Natural Language Toolkit. [atualizado em 2016 April 9; citado em 2016 Jul 4].Disponível em: http://www.nltk.org/.

[12]Zani SL. Álgebra Linear [Internet]. Departamento de Computação e Matemática ICMC - USP. [citado em 2016 Ago 4].Disponível em: http://www.icmc.usp.br/~szani/alglin.pdf.

Apêndice A: Tabela de tagset do NLTK[10], usadas no processamento.

Tag	Significado	Tag	Significado
CC	conjunction, coordinating	PRP$	pronoun, possessive
CD	numeral, cardinal	RB	adverb
DT	determiner	RBR	adverb, comparative
JJ	adjective or numeral, ordinal	RBS	adverb, superlative
JJR	adjective, comparative	RP	particle
JJS	adjective, superlative	VB	verb, base form
NN	noun, common, singular or mass	VBD	verb, past tense
NNP	noun, proper, singular	VBG	verb, present participle or gerund
NNPS	noun, proper, plural	VBN	verb, past participle

Contatos

Gilvan Machado Morais
gilvan.morais@usp.br,

Gabriel Miranda Rubio
gabriel.rubio@usp.br,

Ester E. D. dos Santos
ester.evelyn.santos@usp.br

Alessandra Alaniz Macedo
ale.alaniz@usp.br

Renato de Freitas Bulcão-Neto
rbulcao@ufg.br