UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS
MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

NONILTON ALVES DE SANTANA

A utilização da técnica de filtragem de conteúdo em campos de texto livre para recomendações de diagnósticos.

SÃO LUÍS
2017

NONILTON ALVES DE SANTANA

A utilização da técnica de filtragem de conteúdo em campos de texto livre para recomendações de diagnósticos.

Dissertação apresentada ao Mestrado Profissional de Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação e Sistemas.

Orientador: Prof. Dr. Fernando Jorge Cutrim Demétrio

SÃO LUÍS
2017

AGRADECIMENTOS

Em primeiro lugar agradeço a DEUS pela benção de estar vivo, lúcido, forte e saudável para encarar tamanho desafio. Obrigado a minha linda, adorável, inteligente, paciente e batalhadora esposa pela companhia e incontáveis conselhos ao longo do caminho. Não poderia deixar de agradecer meu avô Manoel José de Santana (in memorian) pelo primeiro investimento em minha carreira, cheguei aqui graças ao senhor. Aos meus pais Raimundo Nonato Neto e Josefa Alves de Santana por dar-me a educação necessária para chegar até aqui, deixo meu muito obrigado, espero um dia retribuir tamanha prova de amor. Ao meu amigo Cleber Augusto pela força no início da caminhada, porque sem ele nada disso teria acontecido, obrigado. Aos professores do programa de mestrado que nos aceitaram em suas disciplinas regulares, mesmo comparecendo a cada 15 dias em função da distância e de não podemos nos mudar para a capital, não consigo descrever tamanha honra, em conhece-los. Obrigado meu irmão Nonailton Alves de Santana por se encarregar ser motorista da minha esposa e filha na minha ausência. Obrigado também ao meu velho amigo Emmanuel Silva Xavier pela parceria de longa data e de valor inestimável.

Não poderia deixar de agradecer aos professores aos professores Fernando Jorge Cutrim Demétrio e Luís Carlos Fonseca pelo apoio e imensurável contribuição nesta jornada.

Gostaria de agradecer a Fundação de Amparo à Pesquisa e Desenvolvimento Científico do Maranhão – FAPEMA pelo financiamento desta pesquisa.

“Mil cairão ao teu lado, dez mil a tua direita, mas tu não serás atingido”

Salmo 91:7

Resumo

O objetivo deste trabalho é propor um sistema de recomendação de diagnóstico médico utilizando a técnica de filtragem de conteúdo. Utilizando a medida de cossenos para recuperação de informação e a medida de pesos TF-IDF (Term Frequency – Inverse Document Frequency) para analise textual, foi possível propor o modelo, dividindo-o em quatro etapas: criação da base de dados de atendimento médicos, criação do mecanismo de buscas por palavras chaves, criação do ranking de itens similares e apresentação da lista de recomendações. Como resultado, a busca utilizando sintomas apresentou resultados satisfatórios e proporcionou liberdade de pesquisa ao usuário. Já a classificação dos prontuários através da similaridade apresentou taxa de acerto de 70% dos diagnósticos submetidos ao sistema.

Palavras-chave:

PEP, Sistema, Saúde, WebService, TF-IDF, Recomendação

ABSTRACT

The objective of this work is to propose a medical diagnostic recommendation system using the content filtering technique. Using the measurement of cosines for information retrieval and the measurement of weights TF-IDF (Frequency Term - Inverse Document Frequency) for textual analysis, it was possible to propose the model, dividing it into four steps: creation of the database of medical care , Creation of the search engine by keywords, creation of ranking of similar items and presentation of the list of recommendations. As a result, the search using symptoms presented satisfactory results and provided freedom of research to the user. On the other hand, the classification of medical records by similarity showed a 70% accuracy of the diagnoses submitted to the system.

Keywords:

PEP, System, Health, WebService, TF-IDF, Recommendation

Lista de figuras

Figura 1 - Filtragem hibrida 16
Figura 2 - Resultado de uma expressão booleana conjuntiva (AND) 18
Figura 3 - Resultado de uma busca booleana disjuntiva OR 19
Figura 4 - Medida dos cossenos para cálculo de similaridade entre documentos 20
Figura 5 - Organograma da pesquisa 24
Figura 6 - Representação dos diagnósticos após formatação dos dados. 26
Figura 7 - Método responsável por recuperar os documentos do corpus. 28
Figura 8 - Classe para tratamento de stopwords 29
Figura 9 - Classe para tratamento de advérbios 29
Figura 10 - Formula para cálculo de pesos 30
Figura 11 - Classe para cálculo de frequência de um termo 30
Figura 12 - Classe para calcular o IDF de um termo. 31
Figura 13 - Matriz de termos e pesos da dengue 31
Figura 14 - Diagrama de classe para persistência da matriz de documentos processados. 32
Figura 15 - Cálculo de pesos dos termos da consulta. 32
Figura 16 - Calculo de pesos dos termos da consulta. 33
Figura 17 - Formula utilizada nesta pesquisa para cálculo de similaridade. 33
Figura 18 - Classe para cálculo de similaridade entre termos da consulta e documentos. 34
Figura 19 - Consulta exemplo com termos genéricos. 34
Figura 20 - Diagrama de componentes representando a arquitetura do sistema 37
Figura 21 - Hipótese diagnostica 38
Figura 22 - Qualidade das recomendações. 48
Figura 23 - Definição do projeto no Firebase 54
Figura 24 - Ferramentas do Firebase 55
Figura 25 - Dependências do Firebase 55
Figura 26 - Diagrama de caso de uso prontuário eletrônico do paciente 57
Figura 27 - Diagrama de classes prontuário eletrônico do paciente 57
Figura 28 - Diagrama de caso de uso modulo administrador 59
Figura 29 - Diagrama de classes modulo administrador 59
Figura 30 - Diagrama de casos de uso modulo de atendimento 60
Figura 31 - Diagrama de classes modulo de atendimento 61
Figura 32 - Classe entidade identificação do médico. 65
Figura 33 - Classe entidade identificação do paciente 65
Figura 34 - Ficha anamnese 66
Figura 35 - Estratégia de persistência de dados classe entidade pacientes. 67
Figura 36- Implementação da classe Firebase. 67
Figura 37 - Menu do protótipo de prontuário eletrônico. 68
Figura 38 - Identificação do médico 69
Figura 39 - Identificação do paciente 69
Figura 40 - Lista de pacientes cadastrados. 70
Figura 41 - Opções do prontuário eletrônico 70
Figura 42 - Parte superior da ficha anamnese 71

Lista de quadros

Quadro 1 - Doenças cadastradas 26
Quadro 2- Fluxo de indexação de um documento. 28
Quadro 3 – Sintomas selecionados aleatoriamente pertencentes ao banco de dados. 40
Quadro 4 - Lista de recomendações para dengue hemorrágica. 41
Quadro 5 - Teste de recomendação com sintomas do Zika Virus. 42
Quadro 6 - Teste de recomendação com sintomas da Febre Chikungunya. 42
Quadro 7 - Teste de recomendação com sintomas da Gripe. 43
Quadro 8 - Teste de recomendação com sintomas da Febre amarela. 44
Quadro 9 - Teste de recomendação com sintomas da Sarampo. 45
Quadro 10 - Teste de recomendação com sintomas da Urticária. 45
Quadro 11 - Teste de recomendação com sintomas da Febre Tifoide. 46
Quadro 12 - Teste de recomendação com sintomas da Pneumonia. 47
Quadro 13 - Teste de recomendação com sintomas da Reumatismo. 47
Quadro 14 - Lista de requisitos funcionais modulo prontuário eletrônico do paciente 56
Quadro 15 - Requisitos funcionais do modulo administrador 58
Quadro 16 - Requisitos funcionais modulo de atendimento de pacientes. 60
Quadro 17 – Identificação do paciente 62
Quadro 18- Identificação do médico 62
Quadro 19 - Login de acesso ao sistema 63
Quadro 20 - Lista de pacientes 63
Quadro 21 - Novo prontuário 63
Quadro 22 - Ficha anamnese 64

Lista de abreviaturas e siglas

SR – Sistemas de recomendação

RI – Recuperação de Informação

SQL – Struct Query Language

SGBD – Sistema gerenciador de banco de dados

Sumário

1. INTRODUÇÃO 11

2. REFERENCIAL TEÓRICO 14

2.1. Sistemas de Recomendação (SR) 14

2.1.1. Filtragem por conteúdo 15

2.1.2. Filtragem Colaborativa 15

2.1.3. Filtragem hibrida. 15

2.2. Aspectos conceituais da recuperação de informação 17

2.2.1. Modelos de RI 17

2.2.2. Modelo Booleano 18

2.2.3. Modelo Vetorial 19

2.2.4. Modelo Probabilístico 21

3. REFERENCIAL METODOLÓGICO 22

3.1. Especificação do modelo 23

3.2. Etapas iniciais 24

3.2.1. Construção de uma base de dados representando os diagnósticos médicos. 24

3.2.2. Mecanismo de busca 27

3.2.2.1. Preparação dos documentos 27

3.2.2.2. Indexação dos termos 28

3.2.3. Classificação de prontuários por analise de conteúdo 32

3.3. Apresentar lista de recomendações 34

3.4. Resumo do modelo 35

4. ARTEFATO – IMPLEMENTAÇÃO DO MODELO 36

4.1. Vantagens 38

4.2. Desvantagens 38

4.3. Testando e validando as recomendações do sistema 39

5. CONSIDERAÇÕES FINAIS 50

6. REFERÊNCIAS 52

Apêndice A - Criando o projeto no Firebase 54

Apêndice B – Requisitos e modelagem do protótipo 56

Apêndice C - Requisitos do protótipo do prontuário eletrônico 62

Implementação dos requisitos 64

1. INTRODUÇÃO

Todas as áreas envolvidas em algum processo de trabalho em vários momentos geram e armazenam informações valiosas para que possam servir de fontes de pesquisas em algum momento, seja no aperfeiçoamento de uma técnica, no desenvolvimento de um novo método ou mesmo na academia. Na saúde, assim como em outras áreas, dado o avanço da tecnologia da informação, realizar este armazenamento já é bastante comum, contudo utilizar estas fontes significa em muitos casos um grande problema, devido os atuais sistemas de gestão utilizados nos hospitais estão voltados para a administração hospitalar, ou seja, estão fortemente direcionados a setores administrativos, financeiros, contábeis, departamento pessoal entre outros.

É sabido que muita informação é coletada de um paciente por meio de um prontuário médico e também que muitos profissionais podem ser envolvidos durante seu atendimento. De acordo com o Conselho Regional de Medicina do Estado de São Paulo (CREMESP) o prontuário eletrônico é um conjunto de documentos padronizados e ordenados, destinados ao registro dos cuidados profissionais prestados ao paciente pelos serviços de saúde pública e privado. Cada procedimento deve ser documentado de forma a facilitar o acompanhamento da evolução clinica deste paciente pela equipe médica. Ainda segundo o CREMESP o objetivo deste documento é catalogar a assistência médica prestada ao paciente, servindo também para ensino e pesquisa e também como instrumento de defesa legal. Percebe-se aqui a relevância de um mecanismo de busca adequado para pesquisa médica, uma vez que estes dados podem duplicar em pouco tempo.

Embora exista no mercado fabricante de software para gestão hospitalar e ofereçam um grande conjunto de ferramentas que dão suporte ao cuidado médico, boa parte deles operam de modo isolado, não oferecendo interfaces de comunicação com outros sistemas ou mesmo um mecanismo inteligente que possa facilitar e/ou mesmo disponibilizar em tempo real uma forma eficiente de consultar informações já armazenadas.

Em sua grande maioria os sistemas de gestão hospitalar oferecem meios de pesquisas através de instruções SQL (Struct Query Language) onde os usuários informam valores que são enviados a sistemas gerenciadores de bancos de dados (SGBD) como Oracle, Microsoft SQL Server, Postgres, Firebird entre outros, afim de encontrar registros que contenham os valores informados. Estas consultas baseiam-se na existência ou não dos termos informados podendo deixar de fora da lista de resultados registros de interesse do usuário.

A reutilização dos dados de um atendimento tem finalidade diversas, dentre elas pode-se destacar o diagnóstico e as condutas médicas utilizadas, a utilização de mecanismo adequado de busca de informação pode ajudar durante os novos atendimentos, agilizando e facilitando o diagnóstico e quais condutas podem ser indicadas ao novo paciente.

Toda a responsabilidade de um pré-diagnostico do paciente depende única e exclusivamente do conhecimento e experiência médica, então de que forma um sistema de recomendação que utiliza filtragem baseada em conteúdo aplicado a uma base de dados de um prontuário eletrônico de paciente poderia auxiliar um médico na definição do diagnóstico do paciente?

1.1. Objetivo geral

Propor um modelo de recomendação de diagnóstico médico através da técnica de filtragem de conteúdo utilizando campos de texto livre de um prontuário eletrônico de paciente.

1.2. Objetivos específicos

- Construir uma base de dados representando os prontuários de atendimento médicos.
- Criar um mecanismo de busca a partir de palavras-chave, resultando em uma lista de diagnósticos médicos classificados pela sua relevância.

- Classificar os prontuários médicos por diagnóstico, através de análise de conteúdo.
- Apresentar a lista de recomendações a partir do ranqueamento formado pela similaridade entre as palavras-chave e prontuário já gravados.

2. REFERENCIAL TEÓRICO

2.1. Sistemas de Recomendação (SR)

O ato de recomendar existe sempre que temos a boa intenção de expressar nossa satisfação ou decepção com algum tipo de serviço, produto, área turística, empresa e até pessoas. Sempre que precisamos utilizamos dessa estratégia para sabermos antecipadamente de alguma informação a respeito de alguma coisa. (WEITZEL; PALAZZO; OLIVEIRA, 2010) destacam que um SR tem o objetivo de auxiliar no fornecimento de sugestões personalizadas de forma automática de acordo com o interesse do usuário. Já (U, 2013) define um SR como uma subclasse de sistemas de filtragem de informação que procuram prever a preferência que um usuário daria para um item. (CAPPELLA; YANG; LEE, 2015; VERBERT K.; MANOUSELIS, 2012) concorda que o objetivo de qualquer sistema de recomendação é estimar a avaliação do usuário de um item que ele não conhece baseando na avaliação de outros itens visto por este usuário. O estudo deste tipo de sistema tornou-se bastante relevante na década de 1990 e mostra-se muito interessante devido ao vasto número de problemáticas e abundancia de aplicações práticas das mesmas. Estes sistemas recolhem informações acerca das preferências de um ou mais usuários sobre um determinado conjunto de itens.

O comercio eletrônico faz uso de recomendação há bastante tempo e sempre que possível, apresenta opiniões de usuários a respeito de determinado produto no qual já demonstrou interesse no passado, ou mesmo elabora lista de itens nos quais você poderia se interessar (VIEIRA; NUNES, 2012; YANG et al., 2016). Um item é um termo utilizado para denominar algo que o sistema recomenda a um usuário, ou seja, pode ser um filme, uma música, livro, hotel e etc. (FADAEE; SOUFIANI; SUNDARAM, 2016; MIGUEL; COSTA, 2016). Serviços como Youtube, Netflix, Amazon sempre nos mostra conteúdo relacionado ao que estamos vendo, ouvido e/ou precisando baseado no seu histórico de navegação (GARCIA; FROZZA, 2013) .

De acordo com (U, 2013; VIEIRA; NUNES, 2012) sistemas de recomendação estão divididos em três categorias, filtragem baseada em conteúdo, filtragem colaborativa e filtragem hibrida.

2.1.1. Filtragem por conteúdo

Este tipo de filtragem utiliza informações anteriores do usuário em relação a um item para recomendar itens similares, por exemplo, recomenda produtos parecidos com aqueles avaliados anteriormente de forma positiva (U, 2013). Neste cenário utiliza principalmente tags, palavras-chaves e descrição dos itens para serem comparados utilizando alguma técnica de recuperação de informação, sendo a indexação de frequência de termos bastante utilizada (CAPPELLA; YANG; LEE, 2015).

Neste tipo de indexação, os itens são descritos detalhadamente e são separados como palavras-chave através de alguma técnica de recuperação de informação, tendo seus termos extraídos com seus respectivos pesos, possibilitando utilizar estas medidas para determinar a similaridade entre os itens (YANG et al., 2016). Em caso de itens serem artigos ou documentos, este processo pode ser ainda mais fácil, pois documentos podem ser considerados similares se compartilharem termos em comum (YANG et al., 2016).

2.1.2. Filtragem Colaborativa

Este tipo de filtragem baseia-se no julgamento de usuários com interesses em comum. São avaliados itens do sistema que permitem descobrir médias para os itens, com isso, o sistema pode descobrir padrões de comportamento e sugerir itens considerados interessantes pelos usuários com gostos similares. A técnica de descoberta automática de relações entre usuário e sua vizinhança consiste em (COSTA; AGUIAR; MAGALHÃES, 2013) :

- Calcular a similaridade do usuário alvo em relação aos outros usuários.
- Selecionar um grupo de usuários com maiores similaridades e considerar a predição.
- Normalizar as avaliações e computar as predições, ponderando as avaliações dos usuários mais similares.

2.1.3. Filtragem hibrida.

De acordo com (RAMOS J. G. A., 2010) este tipo de filtragem é baseado nos métodos filtragem colaborativos e filtragem de conteúdo, reunindo suas principais características levando a uma melhor solução.

(CAZELLA; NUNES; REATEGUI, 2010; FADAEE; SOUFIANI; SUNDARAM, 2016) afirma que a filtragem hibrida procura, combinar os pontos fortes da filtragem colaborativa e filtragem baseada em conteúdo visando criar um sistema que possa melhor atender as necessidades do usuário. A figura 1 apresenta as vantagens envolvidas na filtragem de conteúdo e filtragem colaborativa.

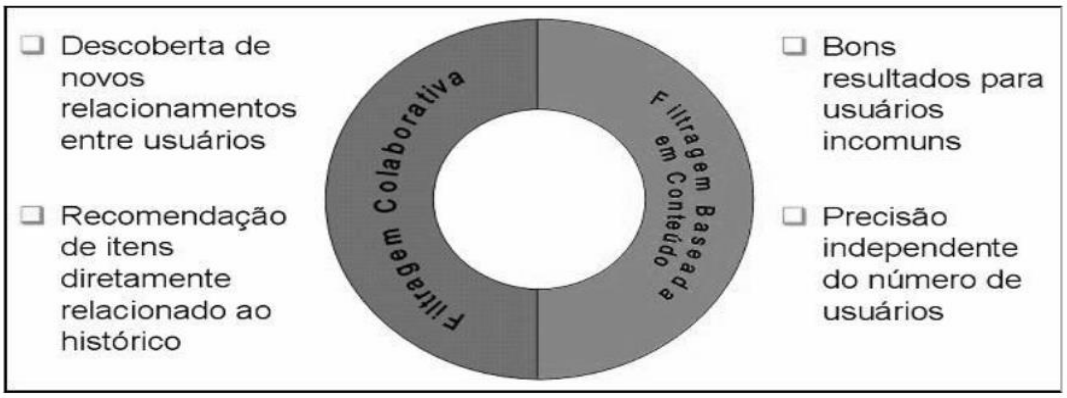

Figura 1 - Filtragem hibrida

Fonte: (CAZELLA; NUNES; REATEGUI, 2010)

A ideia de um sistema de filtragem hibrido é incorporar as melhores características da filtragem colaborativa e filtragem de conteúdo sem herdar as desvantagens

Embora um SR seja bastante comum em outras áreas como comercio eletrônico, entretenimento, notícias entre outros, na medicina este tipo de abordagem é pouco utilizado, sendo aplicado para filtragem de conteúdo de sítios na internet conforme a proposta de (WEITZEL; PALAZZO; OLIVEIRA, 2010) onde sugerem um filtro inteligente através da análise de conteúdo para classificar os sites conforme a qualidade de suas publicações direcionando-as conforme o perfil do usuário.

Observa-se que um SR necessita de outras subáreas da computação como os sistemas de recuperação de informação afim de processar previamente um documento. A seção a seguir descreve os aspectos conceituais e os principais modelos de um sistema de recuperação de informação (RI).

2.2. Aspectos conceituais da recuperação de informação

Todos os dias milhares de pessoas são atendidas em hospitais da rede pública e privada, gerando um enorme volume de dados. Estes registros estão em sua grande maioria armazenados em um sistema estruturado e com suporte a consultas permitindo a recuperação de alguma informação de acordo com a necessidade do usuário. Neste cenário a ciência da recuperação da informação estuda a criação de algoritmos visando recuperar informações a partir de campos de textos livres. De acordo com (BAEZA-YATES; RIBEIRO-NETO, 2013; PAIK, 2013) recuperação de informação trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas web, catálogos online, registros estruturados e semiestruturados, objetos multimídia e etc.

A recuperação de informação (RI) sempre foi um assunto de interesse da humanidade, pois sempre houve a necessidade de preservação e posterior uso do conhecimento registrado. Os livros oferecem há muito tempo um mecanismo simples de RI, o sumário, nele podemos identificar de forma mais eficiente o capitulo ou página de interesse.

2.2.1. Modelos de RI

(BAEZA-YATES; RIBEIRO-NETO, 2013; KIDO; JUNIOR; MORIGUCHI, 2014) afirmam que modelagem em RI é complexo e tem objetivo de produzir função de ranqueamento, produzir escores a documentos em relação a uma consulta. Pode ser dividido em duas fases:

- Concepção de uma coleção de documentos e consultas.
- Definição de uma função de ranqueamento que consulta o grau de similaridade de cada documento em relação a consulta.

(SOUSA; TABOSA, 2015) destaca que todos os sistemas de RI estão baseados nos modelos clássicos de RI, que são:

- Modelo Booleano
- Modelo Vetorial
- Modelo Probabilístico

2.2.2. Modelo Booleano

Neste modelo, documentos são representados por termos indexados, podem ser definidos manualmente ou automaticamente, por meio de algoritmos computacionais. As buscas são formuladas por expressões booleanas constituídas por termos ligados por operadores lógicos (FERNEDA, 2003). Esse modelo é baseado na teoria de conjuntos e na álgebra Booleana. Mostra-se bastante intuitivo e possui semântica precisa (BAEZA-YATES; RIBEIRO-NETO, 2013).

Sua definição diz que os elementos da matriz de termos por documentos são 1 para indicar a presença do termo no documento, ou 0, para indicar sua ausência. As frequências na matriz de termos por documentos são todas binárias. Sendo uma consulta q composta por termos de indexação ligados por três conetivos booleanos: *NOT, AND e OR*. Logo uma consulta é uma expressão Booleana convencional sobre os termos de indexação (BAEZA-YATES; RIBEIRO-NETO, 2013).

Uma expressão conjuntiva de enunciado t_1 AND t_2 recuperará documentos indexados por ambos os termos. Esta operação é a interseção dos conjuntos dos documentos indexados pelo termo t_1 com o conjunto dos documentos indexados pelo termo t_2 conforme figura 2.

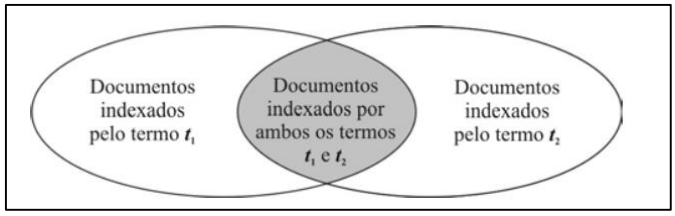

Figura 2 - Resultado de uma expressão booleana conjuntiva (AND)
Fonte: (FERNEDA, 2003)

Para uma expressão disjuntiva t_1 OR t_2 recuperará documentos indexados pelo termo t_1 ou pelo termo t_2, sendo equivalente a união entre os conjuntos dos documentos indexados pelo termo t_1 e o conjunto dos documentos indexados pelo termo t_2, conforme figura 3.

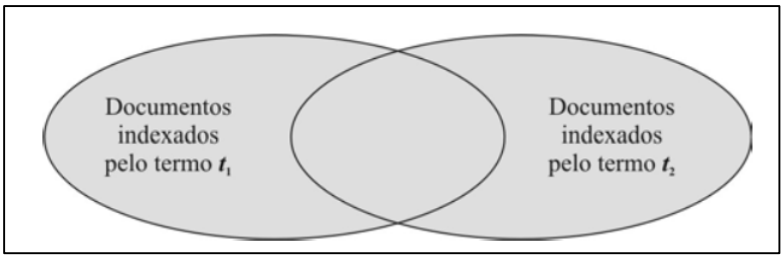

Figura 3 - Resultado de uma busca booleana disjuntiva OR
Fonte: (FERNEDA, 2003).

Por se tratar de consultas representadas por expressões booleanas, várias combinações entre os operadores e os termos de interesse podem ser combinadas afim de satisfazer as necessidades de busca por parte do usuário.

(BAEZA-YATES; RIBEIRO-NETO, 2013) afirma que as principais vantagens do modelo Booleano são o formalismo claro e sua simplicidade. Já (SOUZA, 2006) define como a principal desvantagem deste modelo pouca eficiência devido ao fato de trabalhar de forma binaria, considerando apenas o critério relevante/não relevante e não ordenar resultados de acordo com a consulta. (BAEZA-YATES; RIBEIRO-NETO, 2013) também destacam que consultas Booleanas é inconveniente para a maior parte dos usuários.

2.2.3. Modelo Vetorial

Observando um conjunto de documentos percebemos que alguns termos não acrescentam sentido em relação ao conteúdo do documento, se analisarmos uma coleção com vários documentos, uma palavra que aparece em todos eles, não é muito útil para descrever esta coleção, portanto nada acrescenta sobre quais documentos um usuário estaria interessado. No entanto palavras que apareçam em poucos documentos podem ser bastante interessantes para o usuário (BAEZA-YATES; RIBEIRO-NETO, 2013).

O modelo Vetorial também é conhecido como Modelo Espaço Vetorial e baseia-se no casamento parcial de termos (palavras) por meio da atribuição de pesos não binários para termos da consulta e documento (PAIK, 2013). (MARTHA; DE CAMPOS; SIGULEM, 2010) afirma que este é um dos modelos mais utilizados por representar a informação sob perspectiva matemática e estatística e apresenta bons resultados.

Neste modelo cada documento da coleção é representado por um vetor conforme a expressão $D_j = \{t_1,t_2,t_3,t_4...t_N\}$. Sendo vetor D_j um documento da coleção e $t_1,t_2,t_3,t_4..t_N$ termos que acrescentam valor semântico ao documento e $_j$ o número do documento *D* na coleção. Palavras como preposições, conjunções e advérbios são tratadas como stopwords e devem ser removidas do documento, por nada dizerem sobre o conteúdo do próprio documento (IRACICABA, 2006). A quantidade de aparições de um termo em um documento é contabilizada como sua frequência podendo ser definida por ***TF*** (*Term Frequency*) e pode ser demonstrada pela expressão $\boldsymbol{TF_{ij} = F_{ij}}$.

A frequência de termos pode ser muito interessante para descrever um documento em particular, porém se este termo existir na maioria dos documentos da coleção, este perde sua especifidade e naturalmente pode não representar documentos de interesse na consulta do usuário. Para este problema surge a frequência inversa de documentos (*IDF*) que pode ser representada pela formula $\boldsymbol{IDF = Log\ (N / n_i)}$ sendo *N* o número total de documentos da coleção e n_i o número de documentos em que o termo $_i$ aparece. O peso do termo no documento é calculado pelo formula $\boldsymbol{Wij = TF_{ij} * IDF_{ij}}$ (BAEZA-YATES; RIBEIRO-NETO, 2013).

Em seguida estes pesos são utilizados para calcular o grau de similaridade entre documentos armazenados no sistema e a consulta do usuário para que possam criar um ranking ordenado de resultados através das equações representadas pela figura 5 (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$C_{a,b} = \frac{\sum_{i=1}^{m} (w_{a,i} * w_{b,i})}{\sqrt{\sum_{i=1}^{m} (w_{a,i})^2} \times \sqrt{\sum_{i=1}^{m} (w_{b,i})^2}}$$

Figura 4 - Medida dos cossenos para cálculo de similaridade entre documentos
Fonte: adaptado de (COSTA; AGUIAR; MAGALHÃES, 2013)

As medidas de similaridade entre documentos trabalham com métricas e determinam quando um documento é similar a outro. O uso da medida dos cossenos retorna 0 e 1. Sendo medido o ângulo entre os dois vetores num espaço vetorial. Quanto

mais próximo de 1 for o valor, menor é o ângulo e consequentemente maior à similaridade entre o documento e os termos da consulta (MAIA; SOUZA, 2013).

(BAEZA-YATES; RIBEIRO-NETO, 2013) apresenta algumas observações sobre este modelo, são elas:

- Conjunto ordenado de documentos é retornado, fornecendo uma melhor resposta à consulta.
- Documentos que tem mais termos em comum com a consulta tendem a ter maior similaridade.
- Termos com maiores pesos contribuem mais para o casamento do que os quem tem pesos menores.
- Documentos maiores são favorecidos.
- A similaridade calculada não tem limite superior definido.

2.2.4. Modelo Probabilístico

De acordo com (BAEZA-YATES; RIBEIRO-NETO, 2013) é que existe um conjunto de documentos ideal, dada uma descrição desse conjunto poderíamos recuperar documentos relevantes, sendo a consulta um processo de especificação das propriedades deste conjunto de documentos ideal. O problema deste modelo é que inicialmente não se sabe quais são estas propriedades, forçando uma estimativa inicial destas propriedades. Este modelo reutiliza o resultado da primeira consulta para refinar ainda mais seus resultados (BRITO, 2016) .

3. REFERENCIAL METODOLÓGICO

3.1. Tipo de pesquisa

A pesquisa terá uma natureza **quantitativa** a partir do caráter **descritivo** e **exploratório**. (WAINER, 2007) □ afirma que a pesquisa quantitativa tem suas bases nas ciências naturais observando poucas variáveis e sempre busca obter medidas numéricas. Tem uma visão positivista, ou seja, mesmo diferentes observadores terão o mesmo resultado. (FONSECA, 2002) □ destaca que pesquisa quantitativa é aquela que pode quantificar alguma coisa. Necessita de amostras geralmente grandes e representativas da população e seus resultados são tomados como um retrato real de toda a população-alvo da pesquisa. Recorre a linguagem matemática para descrever causas de um fenômeno, relações entre variáveis e etc. (CHAER; DINIZ; RIBEIRO, 2012)□ afirma que a pesquisa quantitativa procura quantificar os dados fazendo uso de alguma forma de análise estatística.

O caráter descritivo da pesquisa exige do investigador informações sobre o que deseja pesquisar. Este tipo de estudo pretende descrever fatos e fenômenos de determinada realidade (SILVEIRA; CÓRDOVA, 2009) □. Já □(SANTOS, 2008) afirma que pesquisa descritiva descreve uma experiência, uma situação, um fenômeno ou processo nos mínimos detalhes. (FONSECA, 2002)□ concorda que este tipo de pesquisa tem a finalidade de descrever as características de determinada população ou fenômeno, ou o estabelecimento de relação entre variáveis.

O caráter exploratório da pesquisa proporciona ao pesquisar maior familiaridade com o problema em estudo. Seu objetivo é facilitar o entendimento de um problema complexo ou mesmo construir hipóteses mais adequadas □(VIEIRA, 2002) . Já □ (ELETR; GARCIA, 2015) define pesquisa exploratória como a busca pela familiarização dos fenômenos surgidos durante a pesquisa, explorando mais profundamente e com maior precisão. (SILVA; KARKOTLI, 2011)□ entende que a pesquisa exploratória proporciona maior proximidade com o problema visando torná-lo explicito ou definir hipóteses. Também pode aprimorar ideias ou fazer intuições.

3.2. Universo e amostra e critério de seleção

A pesquisa será realizada na cidade de Imperatriz-MA. O universo da pesquisa será representado por uma base de dados contendo 95 doenças com seus respectivos sintomas. De acordo com (GUIMARÃES, 2012; PROVDANOV; FREITAS, 2013) universo da pesquisa representa o conjunto de seres animados ou não que apresenta pelo menos uma característica comum, sendo N o número total de elementos ou população.

Desse universo, serão utilizados 20 conjuntos selecionados baseados nas ocorrências mais frequentes da atenção básica de saúde nos hospitais de acordo com (PIMENTEL et al., 2012). Quanto a amostra, (PROVDANOV; FREITAS, 2013) explica que é uma pequena parte do universo.

A metodologia deste trabalho seguiu as diretrizes do método *Design Science Research*, que segundo (BAX, 2014) é uma estratégia de pesquisa capaz de orientar, tanto a construção do conhecimento, quando aprimorar práticas em sistemas de informação e disciplinas relacionadas ao campo gerencial e tecnológico da ciência da informação. É uma metodologia que encoraja o desenvolvimento de artefatos e/ou protótipos visando compreender problemas do mundo real e propor soluções apropriadas, úteis, fazendo o conhecimento avançar.

Neste trabalho propõe-se a utilização de um modelo de sistema de recomendação utilizando técnicas de filtragem baseada em conteúdo com o objetivo de recomendar diagnósticos médicos.

3.3. Especificação do modelo

Consiste na apresentação da arquitetura do sistema de recomendação a ser aplicado sobre a base de dados médica. Em seguida, apesentará a sua especificação, sendo este dividido em:

- Criar um mecanismo de busca a partir de palavras-chave, resultando em uma lista de diagnósticos médicos classificados pela sua relevância.
- Classificar os prontuários médicos por diagnóstico, através de análise de conteúdo.
- Apresentar a lista de recomendações a partir do ranqueamento formado pela similaridade entre as palavras-chave e prontuário já gravados.

- Construir uma base de dados representando os prontuários de atendimento médicos.

A figura 5 apresenta a pergunta problema e também os objetivos gerais e específicos.

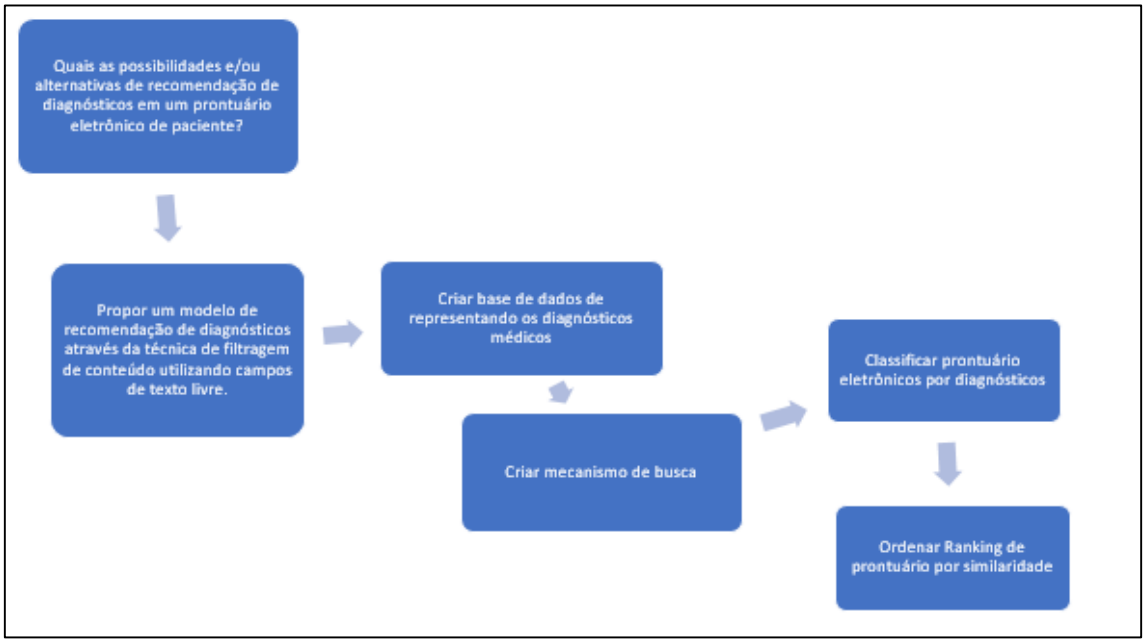

Figura 5 - Organograma da pesquisa

Fonte: Próprio autor.

3.4. Etapas iniciais

3.4.1. Construção de uma base de dados representando os diagnósticos médicos.

Para que o primeiro objetivo especifico seja alcançado foi elaborada uma lista de doenças mais comuns e seus respectivos sintomas. Para esta etapa criou-se uma base de dados experimental contendo um conjunto de 95 enfermidades diferentes, representando prontuários e diagnósticos de pacientes. Contudo, permitiu-se algumas redundâncias totalizando 100 prontuários. O quadro 1 apresenta os diagnósticos cadastrados.

Dengue	Febre amarela	Conjuntivite gastroenterites faringites	Constipação | resfriado	Catapora herpes simplex herpes genital
Gripe	Poliomielite	Hepatite	Caxumba	Sarampo
Rubéola	Raiva (hidrofobia)	Meningite viral	Doença de chagas	Ansiedade
Labirintite	Conjuntivite	Difteria	Reumatismo	Melanoma
Bursite	Alzheimer	Amigdalite	Bulimia	Cistite
Clamídia	Colite	Diverticulite	Distonia	Escoliose
Espondilolistese	Encefalite	Estomatite	Endocardite	Encefalite
Filariose	Foliculite	Gengivite	Feocromocitoma	Gota
Icterícia	Impetigo	Leptospirose	Leucemia	Melanoma
Mononucleose	Narcolepsia	Otite	Parkinson	Progeria
Rabdomiólise	Rosácea	Roséola	Sarcoidose	Seborreia
Sepse	Sífilis	Trombose	Tenossinovite	Uretrite
Urticária	Vaginismo	Varíola	Varicocele	Vulvovaginite
Vitiligo	Zika vírus	Anemia falciforme	Andropausa	Angina
Botulismo	Blefarite	Ciática	Cistite	Câncer de boca
Candidíase	DMRI	Ebola	Febre tifoide	Galactosemia
Gangrena	Gripe H1N1	Hidrocefalia	Hipercalcemia	Hemocromatose
Hemofilia	Hipocondria	Hipocalemia	Hipoparatireoidismo	Hiponatremia

Melasma	Pé de atleta	Periodontite	Queloide	Febre chikungunya
Dengue	Dengue hemorrágica	Pneumonia	Asma	Febre amarela

Quadro 1 - Doenças cadastradas

Fonte: Próprio autor.

Após a etapa de coleta dos dados da primeira etapa, aplicou-se sobre os dados uma estrutura de arquivos no formato JSON (*JavaScript Object Notation*), um formato leve de arquivos utilizado para transferência de dados entre aplicações heterogêneas (SALVADORI, 2014). Cada registro coletado foi posicionado na estrutura "chave": "valor", sendo os sintomas apenas separados por "," (virgula). A figura 6 apresenta o resultado da formatação dos dados.

```
[
    {
        "diagnostico":"dengue",
        "sintomas": [
                        "febre alta",
                        "dores no corpo (Abdominal)",
                        "erupções avermelhadas",
                        "Hemorragias nos casos graves"
                    ]
    },
    {
        "diagnostico":"febre amarela",
        "sintomas": [
                        "febre",
                        "dores",
                        "irritação",
                        "pulso baixo",
                        "enjôo",
                        "pessoa amarelada",
                        "complicações no fígado",
                        "complicações no rins",
                        "complicações no hemorragias"
                    ]
    },
    {
        "diagnostico":"conjuntivite gastroenterites faringites",
        "sintomas": [
                        "inflamação local",
                        "náuseas",
                        "vômitos",
                        "irritação"
                    ]
    },
```

Figura 6 - Representação dos diagnósticos após formatação dos dados.

Fonte: Próprio autor.

Após a formatação dos dados que representam os prontuário e seus respectivos diagnósticos o mesmo foi persistido em um banco de dados na internet, para que posteriormente possa ser indexado pelo algoritmo de cálculo de pesos.

3.4.2. Mecanismo de busca

Esta etapa consiste na criação de um método que permita ao usuário informar um conjunto de palavras-chaves que serão submetidas ao sistema e consequentemente uma lista de resultados deverá ser retornada ao usuário. No entanto os documentos da coleção podem não estar no formato esperado pelo sistema, devendo então passar por um processo de homogeneização para que o sistema execute o correto cálculo de pesos de seus termos.

3.4.2.1. Preparação dos documentos

Nesta parte do processo o documento é carregado da base de dados e nesta pesquisa cada documento é representado por um único diagnostico armazenado no arquivo JSON, ou seja, cada diagnostico é um documento da coleção. Após serem identificados as etapas do quadro 2 são aplicadas a cada documento.

Etapa	Descrição

Analise do documento	Consiste em preparar o documento para indexação, está dividia em 3 fases:
	• Extração
	  ○ Consiste em transformar, formatar em codificação universal como UTF-8 e validar o documento.
	• Normalização
	  ○ Remoção de sinais e caracteres especiais, formatadores e pontuação em geral
	• Carga
	  ○ Estabelece um novo armazenamento estruturado, confiável e de rápido acesso

	a massa de dados homogeneizada e preparada para indexação
Remoção das Stop-Words	Consiste na remoção de palavras com valor semântico desconsiderável. Estas palavras são especificas para cada idioma e estão dispostas em arquivos denominados stop-list.
Stemming do documento	Esta operação acontece após a remoção das stop-words do documento e trata-se da extração dos radicais semânticos de cada palavra. A raiz de uma palavra consiste na remoção dos sufixos e prefixos, apresentando seu valor essencial.

Quadro 2- Fluxo de indexação de um documento.

Fonte: Próprio autor.

Uma classe foi criada com a responsabilidade de recuperar a lista de documentos da coleção do banco de dados. A figura 7 apresenta sua implementação.

```php
private function getFirebaseDocuments($fileJson) {
    $fb = new Firebase();

    $a = (array) json_decode( $fb->get($fileJson) , true);
    $this->totalDocumentosColecao = count($a);


    foreach ($a as $k=>$v)
    {
        $str = null;

        $doc = new TDocumento();
        $doc->setUid($v['uid']);

        foreach ($v['anamnese'] as $c => $b) {
            $str = $b."\n";
            $doc->setTermos($this->getListaTermos(array_count_values($this->normalize->processa($str))));
            $doc->setTotalTermos(count($doc->getTermos()));

        }

        foreach ($v['hipoteses'] as $c => $b){
            if ($c == "hipotese")
                $doc->setDiagnostico($b);
        }

        $this->documentos[] = $doc;

    }

   // echo json_encode($this->documentos);
    $this->processa();
}
```

Figura 7 - Método responsável por recuperar os documentos do corpus.

Fonte: Próprio autor.

3.4.2.2. Indexação dos termos

No processo de preparação dos documentos é importante utilizar uma lista pré-definida de *stopwords* e uma outra lista de advérbios. Estas duas listas tem a função de

remover do documento palavras que não acrescentam valor semântico, ou seja nada dizem em relação ao próprio texto. Segundo (BRITO, 2016) são exemplos de *stopwords* palavras como “ambas”, “ambos”, “ano”, “anos”, “as”, “a”, “do”, “para” entre outras. Nesta pesquisa utilizou-se a lista de *stopwords* disponibilizada pelo Google. A lista de todos os termos da lista de *stopwords* e lista de advérbios é apresentado no apêndice A. As figuras 8 e 9 apresentam os algoritmos responsáveis por tratar stopwords e advérbios.

```php
class TStopWord {
    //put your code here
    private $stopwords;

    public function TStopWord(){
        $this->stopwords = file("resources/stopwords");

        for($i=0;$i<sizeof($this->stopwords);$i++){
            $this->stopwords[$i] = trim(strtolower($this->stopwords[$i]));
        }
    }

    public function getStopWords(){
        return $this->stopwords;
    }

    public function removeStopWords($content){
        $termos = explode(" ", $content);
        $novosTermos = array();
        for ($i = 0; $i < sizeof($termos); $i++) {
            $termos[$i] = strtolower(trim($termos[$i]));

            if (!in_array($termos[$i], $this->stopwords)){
                $novosTermos[] = $termos[$i];
            }
        }

        return $novosTermos;

    }
}
```

Figura 8 - Classe para tratamento de stopwords.

Fonte: Próprio autor.

```php
class TAdverbios {

    //put your code here
    private $adverbios;

    public function __construct() {
        $this->adverbios = file("resources/adverbios");
        for ($i = 0; $i < sizeof($this->adverbios); $i++) {
            $this->adverbios[$i] = trim(strtolower($this->adverbios[$i]));
        }
    }

    public function getAdberbios() {
        return $this->adverbios;
    }

    public function removeAdverbios($content) {
        $novosTermos = array();
        foreach ($content as $line) {
            if (!in_array($line, $this->adverbios)) {
                $novosTermos[] = $line;
            }
        }

        return $novosTermos;
    }

}
```

Figura 9 - Classe para tratamento de advérbios

Fonte: Próprio autor.

Em seguida a preparação dos documentos cada documento é processado, nesta pesquisa o algoritmo utilizado para o cálculo de pesos TF-IDF (*Term Frequency – Inverse Documento Frequency*) na sua forma clássica por apresentar bons resultados de acordo

com (BAEZA-YATES; RIBEIRO-NETO, 2013). A figura 10 apresenta a equação do modelo *TF-IDF*.

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

Figura 10 - Formula para cálculo de pesos
Fonte: (BAEZA-YATES; RIBEIRO-NETO, 2013)

Neste modelo cada documento da coleção é representado por um vetor conforme a expressão $D_j = \{t_1,t_2,t_3,t_4...t_N\}$. Sendo vetor D_j um documento da coleção e $t_1,t_2,t_3,t_4..t_N$ termos que acrescentam valor semântico ao documento e $_j$ o número do documento D na coleção.

Para determinar o valor *TF (Frequency Term)* de um termo no documento é aplicado a formula $TF_{ij} = Frequência_{ij} / máxima\ Frequência_{ij}$. A figura 11 apresenta a classe para cálculo de frequência de termos.

```
class TFrequencia {

    private $termosFrequencia;
    private $termoMaiorFrequencia;
    private $arrayTermos;

    function getTermoMaiorFrequencia() {...}

    function getArrayTermos() {...}

    function getTermosFrequencia() {...}

    /** Construtora Frequencia ...*/
    public function __construct($termos) {
        $this->convert($termos); //convert o array de objeto TTermos para arrayString;
        $this->termosFrequencia = array_count_values($this->arrayTermos);
        $this->termoMaiorFrequencia = max($this->termosFrequencia);

    }

    /** Convert um array de objeto para um array de strings ...*/
    private function convert($termos) {
        $this->arrayTermos = array();
        for ($i = 0; $i < sizeof($termos); $i++) {
            $this->arrayTermos[] = $termos[$i]->getFrequencia();
        }
    }

}
```

Method 'getFrequencia' not found in more... (Ctrl+F1)

Figura 11 - Classe para cálculo de frequência de um termo.
Fonte: Próprio autor.

Já o IDF determina-se através da formula: $IDF_{ij} = log\ (N / df_i)$, sendo:

- N – O número total de documentos da coleção
- dfi – O número total de vezes que este termo apareceu na coleção de documentos.

A figura 12 apresenta a implementação da classe responsável por calcular o IDF de um termo.

```php
class TCalculaIDF {

    /** Calcula o IDF ...*/
    public function calculaIDF($N, $n) {
        if ($n <= 0) {
            return 0;
        } else {
            return log($N / $n);
        }
    }

}
```

Figura 12 - Classe para calcular o IDF de um termo.

Fonte: Próprio autor.

Após a aplicação da formula sobre os documentos da coleção, uma matriz vetorial contendo documentos, termos e seus respectivos pesos é gerada. Esta matriz vetorial é que deverá ser usada para determinar a distância semântica durante as buscas do usuário. A figura 13 apresenta a matriz de termos referente ao diagnostico com os sintomas da dengue.

Termo	Frequencia	IDF	Peso	QtColeção
febre	0.1	0.96601416722659	0.096601416722659	153
alta	0.1	1.9534008207845	0.19534008207845	57
dores	0.1	1.6526466667653	0.16526466667653	77
corpo	0.1	2.3855341759748	0.23855341759748	37
abdominal	0.1	3.1632387445628	0.31632387445628	17
erupcoes	0.1	3.1632387445628	0.31632387445628	17
avermelhadas	0.1	3.4315027311575	0.34315027311575	13
hemorragias	0.1	3.7992275112828	0.37992275112828	9

Figura 13 - Matriz de termos e pesos da dengue

Fonte: Próprio autor.

Observe a figura 13, o parâmetro *QtColeção*, que nesta imagem representa quantas vezes aquele termo apareceu na coleção de documentos do corpus. Depois de gerado a matriz vetorial de documentos a mesma deve ser persistida o banco de dados, pois a etapa de preparação e indexação dos documentos do corpus é uma atividade que exige alto poder de processamento e deve ser feita previamente. A figura 14 apresenta o diagrama de classes que representa a matriz de documentos.

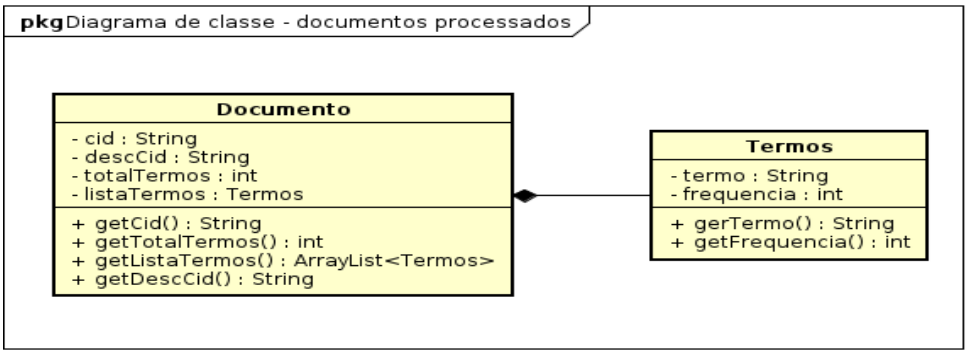

Figura 14 - Diagrama de classe para persistência da matriz de documentos processados.

Fonte: Próprio autor.

3.4.3. Classificação de prontuários por analise de conteúdo

A classificação será feita mediante a informação de palavras-chave durante o atendimento médico e também pode ser entendida com uma consulta do usuário, (MARTHA; DE CAMPOS; SIGULEM, 2010) definem que a consulta é um processo de recuperação e consiste em comparar os termos da pesquisa com os termos do índice e retornar os documentos relevantes à pesquisa ordenados por um critério especificado. É responsabilidade do usuário informar os termos ou palavras-chave de seu interesse visando encontrar no conjunto de resultados um documento do seu interesse. Termos muito genéricos ou de pouca relevância tendem a deteriorar os resultados da consulta. A estratégia de cálculo de similaridade utilizada nesta pesquisa é a medida dos cossenos.

Para que os termos da consulta sejam utilizados no cálculo de similaridade é necessário extrair o peso dos termos da consulta. Nesta pesquisa foi utilizada a formula conforme a figura 15.

$$w_{i,j} = \left(0.5 + \frac{0.5\ freq_{i,q}}{max_l\ freq_{l,q}}\right) X\ \log\frac{N}{n_i}$$

Figura 15 - Cálculo de pesos dos termos da consulta.

Fonte: (BAEZA-YATES; RIBEIRO-NETO, 2013)

Para implementação da formula da figura 15, criou-se dois métodos dentro da classe denominada *TConsulta.* Esta por sua vez é responsável por executar todo o fluxo de execução da consulta por documentos relevantes aos termos informados. A figura 16 apresenta a implementação da formula da figura 16.

```
public function TConsulta($dadosConsulta) {
    $normalize = new TNormalize();
    $this->termos = $this->termosConsulta($normalize->processa($dadosConsulta));
    $this->processa = new TProcessa("prontuarios.json");
    $this->calcularPeso($this->termos);
    $this->calculaSimilaridade();
    $this->imprimir();
}

private function calcularPeso($termos) {
    $maxFreqTermo = max($this->termosFrequencia);

    for ($i = 0; $i < sizeof($termos); $i++) {
        $ocorrencias = $this->processa->totalOcorrenciasColecao($termos[$i]->getTermo());
        if ($ocorrencias == 0) {
            $this->pesos[] = 0;
        } else {
            $this->pesos[] = (0.5 + (0.5 * $this->termosFrequencia[$termos[$i]->getTermo()] / $maxFreqTermo)) * log($this->processa->getTotalDocumentosColecao() / $ocorrencias)
        }
    }
}
```

Figura 16 - Calculo de pesos dos termos da consulta.
Fonte: Próprio autor.

Após a geração dos pesos dos termos da consulta finalmente é utilizada a medida dos cossenos para que a similaridade possa ser determinada e um conjunto de resultados possa ser apresentado para o usuário. A figura 17 apresenta a formula utilizada nesta pesquisa.

$$C_{a,b} = \frac{\sum_{i=1}^{m} (w_{a,i} * w_{b,i})}{\sqrt{\sum_{i=1}^{m} (w_{a,i})^2} \times \sqrt{\sum_{i=1}^{m} (w_{b,i})^2}}$$

Figura 17 - Formula utilizada nesta pesquisa para cálculo de similaridade.
Fonte: (COSTA; AGUIAR; MAGALHÃES, 2013)

No modelo de sistema de recomendação de diagnostico proposto nesta pesquisa o médico representa o usuário, que durante o atendimento de um novo paciente informa os dados do atendimento em um campo de texto livre, o sistema encarrega-se de enviar os termos informados pelo médico representando os termos da consulta para o algoritmo de indexação de modo não intrusivo, o sistema por sua vez, aplica o cálculo de similaridade entre todos os documentos processados e apresenta uma lista de documentos

similares. A figura 18 apresenta a classe responsável por determinar a similaridade entre os termos da consulta e os termos dos documentos.

```php
class TSimilaridade {

    //put your code here

    private function somaDosQuadrados($d) {
        $soma = 0;
        for ($i = 0; $i < sizeof($d); $i++) {
            $soma += pow($d[$i], 2);
        }

        return sqrt($soma);
    }

    public function similaridade($pesosConsulta, $documento) {
        $soma = 0;
        $QtTermosConsulta = sizeof($pesosConsulta);
        $pesoTermosDoc = array();

        $termosDocumento = $documento->getTermos();
        for ($i = 0; $i < sizeof($termosDocumento); $i++) {
            if ($i > $QtTermosConsulta - 1) {
                $soma += 0;
            } else {
                $pesoTermosDoc[] = $termosDocumento[$i]->getPeso();
                $soma += $pesosConsulta[$i] * $termosDocumento[$i]->getPeso();
            }
        }

        if ($soma == 0)
            return 0;
        else
            return $soma / ($this->somaDosQuadrados($pesosConsulta) * $this->somaDosQuadrados($pesoTermosDoc));
    }

}
```

Figura 18 - Classe para cálculo de similaridade entre termos da consulta e documentos.

Fonte: Próprio autor.

Em seguida a figura 19 apresenta uma lista reduzida de resultados não ordenados contendo documentos similares aos termos de uma consulta de exemplo composta pelos termos “dor de cabeça, dor nos olhos, dor no corpo”.

```
diagnostico: dengue similaridade: 0.96214746108861
diagnostico: febre amarela similaridade: 0.99296896953844
diagnostico: conjuntivite gastroenterites faringites similaridade: 0.88923396580168
diagnostico: constipação | resfriado similaridade: 0.96065966128391
diagnostico: catapora herpes simplex herpes genital similaridade: 0.91258665524324
diagnostico: gripe similaridade: 0.92379149527743
diagnostico: poliomielite similaridade: 0.92379149527743
diagnostico: hepatite similaridade: 0.9461759167811
diagnostico: caxumba similaridade: 0.96027316717634
diagnostico: sarampo similaridade: 0.979154424797
diagnostico: rubéola similaridade: 0.96274466786442
diagnostico: raiva (hidrofobia) similaridade: 0.97708645165138
diagnostico: meningite viral similaridade: 0.8753925907376
diagnostico: doença de chagas similaridade: 0.93177240811476
```

Figura 19 - Consulta exemplo com termos genéricos.

Fonte: Próprio autor.

3.5. Apresentar lista de recomendações

A lista de recomendações representa o ranking ordenado de resultados que será apresentado ao médico durante o atendimento. Este resultado é formado pela similaridade encontrada entre os termos informados durante o atendimento e os termos dos documentos indexados previamente. Neste cenário cada documento é definido como

similar se o valor retorna pela classe de cálculo de similaridade apresentar valores entre 0 e 1. Quando um documento apresentar valor de similaridade próximo de 1 significa que este documento é muito similar aos termos informados no atendimento que representa a operação de consulta. A figura 19 apresentou uma lista de resultados com valores interessantes no que diz respeito à similaridade entre os termos, no entanto percebe-se que não existe uma ordem ou ranqueamento definido para os resultados.

Nesta etapa da pesquisa optou-se por utilizar o valor da similaridade encontrada para ordenação dos resultados, para representar o diagnóstico mais adequado para utilização pelo médico, sendo este livre para determinar qualquer outro diagnóstico, o sistema apenas atuará com a recomendação apresentando as 10 melhores opções, baseado em dados anteriores.

3.6. Resumo do modelo

O sistema de recomendação proposto nesta pesquisa tem a finalidade de oferecer sugestões de diagnósticos médicos através da análise de similaridade entre um novo atendimento e os atendimentos anteriores. Para que esta pesquisa apresente os resultados esperados, foi elaborado um protótipo em escala reduzida afim de permitir a inserção dos dados que representam os diagnósticos já realizados pelos médicos. Após esta fase, iniciou-se o processo de definição da estratégia de indexação, persistência e posterior recuperação das informações do banco de dados através da medida de similaridade utilizando a formula dos cossenos. Em seguida foi definido a estratégia de ranqueamento dos resultados obedecendo a princípio o valor da similaridade entre os termos informados pelo usuário e os registros já indexados previamente. Só então foi exibida a lista de resultados para que possa ser utilizada pelo médico durante um novo atendimento.

O médico é livre para utilizar qualquer hipótese diagnostica, o papel do sistema proposto nesta pesquisa é apenas oferecer de forma rápida uma lista de sugestões de acordo com atendimentos anteriores. Deste modo, também oferece uma forma rápida e eficiente de reaproveitar as informações na base de dados dos sistemas hospitalares.

4. ARTEFATO – IMPLEMENTAÇÃO DO MODELO

Para a implementação da arquitetura do modelo de sistemas proposto, foi necessário a construção de um protótipo de prontuário eletrônico em escala reduzida para dar suporte a operação de construção da base de dados que representa os atendimentos médicos. A definição de prontuário eletrônico de paciente (PEP) pode ser entendida como um registro que reside em um sistema especificamente desenhado para apoiar os usuários, fornecendo acesso a um completo conjunto de dados, sistemas de avisos e alertas, sistemas de apoio a decisão e outros (MARIN, 2010). (PEREZ; ZWICKER, 2010) define prontuário eletrônico do paciente como um processo que incorpora registros de um paciente em um sistema informatizado, com o objetivo de gerar informações para diagnósticos médico e documentação de consultas. (PETERS et al., 2010) esclarece que prontuários eletrônicos de paciente são uma fonte importante de informação e deve ser elaborado de forma séria, obedecendo a critérios e linguagem adequada que permita comunicação fácil entre profissionais da área.

Os requisitos e diagramas pertencentes a etapa de construção do protótipo estão disponibilizados no apêndice A.

A tarefa de indexação de termos, dever ser realizada previamente pelo algoritmo para evitar sobrecarga computacional desnecessária. Contudo foi necessário construir uma camada de integração de software capaz de prover informações de modo transparente entre todos os módulos existentes do sistema. Mesmo não sendo um dos objetivos específicos optou-se por utilizar um webservice acessado através da internet para prover capacidade de integração com sistemas de terceiros no futuro. Este tipo de software é denominado webservice REST (*Representational State Transfer*) e está apoiado sobre o protocolo de comunicação HTTP sendo independente de plataformas, oferecendo endereço único para transferência de recursos. Neste modulo permite a integração da computação móvel com a computação em nuvem. Esta camada habilita o acesso ubíquo e pervasivo, um ambiente escalável, distribuído e compartilhado, podendo ser rapidamente provido e liberado com o mínimo de esforço gerencial (COSTA, 2015). A figura 20 apresenta o conceito de arquitetura utilizado no protótipo.

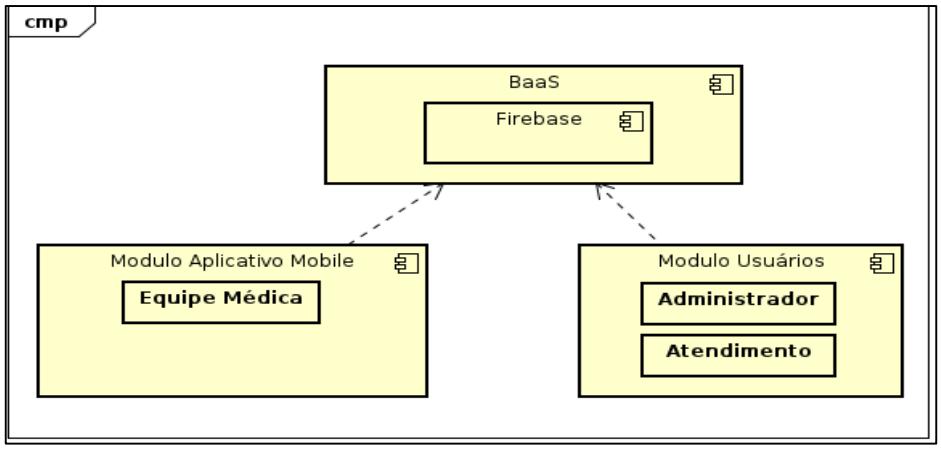

Figura 20 - Diagrama de componentes representando a arquitetura do sistema
Fonte: Próprio autor.

Utilizou-se uma abordagem relativamente nova que utiliza uma nuvem computacional como serviço é o *Mobile Backend as a Service* (*MbaaS*) este tipo de arquitetura oferece serviços de processamento, *storage,* notificações entre outros de forma escalar (COSTA, 2015). Uma boa alternativa para agilizar e reduzir custos durante a construção de sistemas complexos. Várias plataformas foram analisadas e a escolhida foi a plataforma *Firebase* por oferecer um plano *free* com boas configurações para a etapa de desenvolvimento.

O serviço de recomendação atua em conjunto com a camada de consulta, sendo este responsável por submeter as anotações do médico durante o atendimento para o processamento e apresentar a lista de resultados ranqueados para utilização do médico. O médico por sua vez poderá receber as a lista de recomendações na tela de hipótese diagnostica, presente do no protótipo desenvolvido, conforme figura 21.

Figura 21 - Hipótese diagnostica
Fonte: Próprio Autor.

A hipótese diagnostica resultante do processamento de busca por conteúdo através da similaridade é apenas uma sugestão para o médico, não sendo caracterizado como uma obrigação pelo seu uso, contudo apresenta uma forma interessante de associar os novos atendimentos a informações já gravadas no banco de dados, desta forma cria-se uma interface rápida e eficiente para que estas informações possam ser reaproveitas de forma muito mais pratica.

4.1. Vantagens

- Utilizar informações puramente textuais e sem nenhuma estrutura para obter diagnósticos semelhantes.
- A possibilidade de recomendações para definição de diagnósticos.
- Integração com sistemas de terceiros através de API RESTFULL.
- Compartilhamento de informações entre médicos, pacientes e demais interessados.
- Utilização do serviço através da internet.

4.2. Desvantagens

- Necessidade da internet para consumir a API RESTFULL
- Erros de digitação, sinônimos ou jargões médicos podem deteriorar os resultados do algoritmo.
- Poucos registros diminuem a qualidade dos resultados das recomendações.

4.3. Testando e validando as recomendações do sistema

Os testes do sistema foram executados a partir da seleção aleatória de 10 diagnósticos cadastrados previamente na base de dados de testes. O quadro 3 apresenta os diagnósticos selecionados.

Diagnóstico	Sintomas
Dengue hemorrágica	Febre súbita e alta acima de 40, dor atrás dos olhos, falta de apetite e paladar, dos nos ossos e nas articulações, fortes dores de cabeça, manchas vermelhas na pele, náuseas e vômitos intenso, moleza e cansaço, dificuldade de respirar, perda de consciência, confusão mental, agitação e insônia, sangramento na boca nas gengivas e no nariz, boca seca, muita sede, fortes dores abdominais, pele pálida, fria e úmida
Zika Virus	Dor de cabeça, febre baixa, dores leves nas articulações, manchas vermelhas na pele, coceira e vermelhidão nos olhos. Outros sintomas menos frequentes são inchaço no corpo, dor de garganta, tosse e vômitos. No geral, a evolução da doença é benigna e os sintomas desaparecem espontaneamente após 3 a 7 dias. No entanto, a dor nas articulações pode persistir por aproximadamente um mês
Chikungunya	Febre alta, dores intensas nas articulações dos pés, mãos e dedos, tornozelos e pulsos, dor de cabeça, dores nos músculos e manchas vermelhas na pele
Gripe	Febre alta, tosse, dor de cabeça, dores musculares, falta de ar, espirros, dor na garganta, fraqueza, coriza, congestão nasal.

Febre amarela	Febre, dores, irritação, pulso baixo, enjoo, pessoa amarelada, complicações no fígado, complicações nos rins, complicações nas hemorragias
Sarampo	Febre alta, tosse, coriza, manchas avermelhadas pelo corpo, diarreia, otite, pneumonia, encefalite
Urticária	Superfície na pele vermelhos e salientes, coçam intensamente na pele
Febre tifoide	Febre alta, dor de cabeça, mal-estar geral, falta de apetite, retardamento do ritmo cardíaco, aumento do volume do baço, manchas rosadas no tronco, prisão de ventre, diarreia, tosse seca
Pneumonia	Febre alta, Tosse, Dor no tórax, Alterações da pressão arterial, Confusão mental, Mal-estar generalizado, Falta de ar, Secreção de muco purulento de cor amarelada ou esverdeada, Toxemia (danos provocados pelas toxinas carregadas pelo sangue), Prostração (fraqueza)
Reumatismo	Dores nas articulações, principalmente por mais de seis semanas, vermelhidão, calor, inchaço nas articulações, dificuldade para movimentar as articulações ao acordar, dores ao esticar os braços sobre a cabeça, dores ao elevar os ombros até tocar o pescoço

Quadro 3 – Sintomas selecionados aleatoriamente pertencentes ao banco de dados.
Fonte: Próprio autor.

4.2.1 Recomendações para Dengue Hemorrágica

Para o primeiro teste, o diagnostico pré-definido foi a *Dengue Hemorrágica*, utilizou-se os seguintes termos representando a queixa do paciente: *"Febre súbita e alta acima de 40, dor atrás dos olhos, dos nos ossos e nas articulações, fortes dores de cabeça, manchas vermelhas na pele, náuseas e vômitos intenso, moleza e cansaço, dificuldade de respirar, perda de consciência, confusão mental, agitação e insônia, sangramento na boca nas gengivas e no nariz, muita sede, fortes dores abdominais"*. O quadro 4 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.92373092422518	DMRI
0.91357113372885	Zika Vírus
0.89773802863349	Parkinson
0.89230763875844	Hipocondria
0.88174080982724	Dengue hemorrágica
0.82821253776435	Asma
0.79038762578422	Vitiligo
0.76866284634689	Leucemia
0.76286123549829	Hidrocefalia
0.76151903758742	Endocardite

Quadro 4 - Lista de recomendações para dengue hemorrágica.

Fonte: Próprio autor.

No teste com os sintomas da dengue hemorrágica foram informados e apesar de que o diagnóstico esperado foi apresentado na 5ª (quinta posição), os termos informados no primeiro diagnóstico do sistema são genéricos e não descrevem muito bem este prontuário em relação aos prontuários da coleção, por este motivo percebe-se a deterioração dos resultados.

4.2.2 Recomendações para Zika Virus

Para o segundo teste, o diagnostico pré-definido foi o *Zika Virus*, utilizou-se os seguintes termos representando a queixa do paciente: "*Dor de cabeça, febre baixa, dores leves nas articulações, manchas vermelhas na pele, coceira e vermelhidão nos olhos. Outros sintomas menos frequentes são inchaço no corpo, dor de garganta, tosse e vômitos. No geral, a evolução da doença é benigna e os sintomas desaparecem espontaneamente após 3 a 7 dias. No entanto, a dor nas articulações pode persistir por aproximadamente um mês*". O quadro 5 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.98768742933078	Zika Virus
0.90950936466615	Hipocondria
0.89653596897562	DMRI

0.88630556382337	Parkinson
0.87927452259001	Dengue hemorrágica
0.83068199955195	Asma
0.74990256422406	Vitiligo
0.74937268756057	Leucemia
0.73262124421253	Endocardite
0.72250856016646	Hipoparatireoidismo

Quadro 5 - Teste de recomendação com sintomas do Zika Virus.
Fonte: Próprio Autor.

Como esperado o diagnóstico mais bem posicionado foi o *Zika Virus*, contudo percebe-se que mesmo informando todos os sintomas de um determinado prontuário, não é garantido que o mesmo figure entre as primeiras posições da lista de resultados.

4.2.3 Recomendações para Febre Chikungunya

Para o terceiro teste, o diagnóstico pré-definido foi o da *Febre Chikungunya*, utilizou-se os seguintes termos representando a queixa do paciente: "*Febre alta, dores intensas nas articulações, dor de cabeça, dores nos músculos e manchas vermelhas*". O quadro 6 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.94105042504011	Raiva (hidrofobia)
0.93819620236787	Hipocondria
0.93304258486656	Febre chikungunya
0.9209074081749	DMRI
0.91482221194723	Varicocele
0.91347087921938	Roséola
0.91115322933589	Estomatite
0.91056664676546	Endocardite
0.90560474482401	Rosácea
0.90433856456518	Caxumba

Quadro 6 - Teste de recomendação com sintomas da Febre Chikungunya.

Fonte: Próprio Autor.

Neste cenário, mesmo sendo apresentada na terceira posição, a utilização de termos presentes em muitos documentos causou a queda do diagnóstico de interesse na lista de resultados.

4.2.4 Recomendações para Gripe

Para o quarto teste, o diagnóstico pré-definido foi a Gripe, utilizou-se os seguintes termos representando a queixa do paciente: "*Febre alta, tosse, dor de cabeça, dores musculares, falta de ar, espirros, dor na garganta, fraqueza, congestão nasal*". O quadro 7 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.98939550752921	Gripe H1N1
0.97976185167856	Asma
0.96278881065566	Pneumonia
0.93838272457561	Poliomielite
0.93582202914283	Varicocele
0.93465746560358	DMRI
0.93257528644542	Ansiedade
0.93178652230957	Gengivite
0.92290387105008	Encefalite
0.91855858895279	Hipercalcêmica

Quadro 7 - Teste de recomendação com sintomas da Gripe.

Fonte: Próprio Autor.

No teste apresentado no quadro 6, o diagnóstico de interesse ficou fora da lista de resultados, mesmo apresentando *0.9032585695928* de similaridade junto aos termos informados na consulta. Tal fato se dá devido os sintomas da gripe serem muito comuns em relação a outros diagnósticos utilizados na validação do sistema.

4.2.5 Recomendações para Febre Amarela

Para o quinto teste, o diagnóstico pré-definido foi a Febre Amarela, utilizou-se os seguintes termos representando a queixa do paciente: "*Febre, dores, irritação, pulso baixo, enjoo, pessoa amarelada, complicações no fígado, nos rins, e hemorragias*". O quadro 8 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.98350591629966	DMRI
0.96749812426752	Raiva
0.95812171153115	Gangrena
0.95107484103601	Constipação
0.95097620942221	Varicocele
0.9507659658147	Sarcoidose
0.94717675090407	Hemocromatose
0.9464641993411	Sarampo
0.94041752999239	Alzheimer
0.94024719455055	Leucemia

Quadro 8 - Teste de recomendação com sintomas da Febre amarela.
Fonte: Próprio Autor.

Mais uma vez a falta de termos específicos deteriorou a qualidade das recomendações, em função do generalismo com que foi descrito as queixas do paciente.

4.2.6 Recomendações para Sarampo

Para o sexto teste, o diagnóstico pré-definido foi o *Sarampo*, utilizou-se os seguintes termos representando a queixa do paciente: "*Febre alta, tosse, coriza, manchas avermelhadas, diarreia, otite, pneumonia, encefalite*". O quadro 9 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.98245075366334	Febre chikungunya
0.98075412409922	Sarampo
0.97731782889731	Dengue
0.96392864032761	DMRI

0.9512924555291	Varicocele
0.95097961370295	Andropausa
0.9435196206008	Hemocromatose
0.94101365160148	Gripe H1N1
0.93541312651265	Rosácea
0.93442088302775	Hiponatremia

Quadro 9 - Teste de recomendação com sintomas da Sarampo.

Fonte: Próprio Autor.

No teste apresentado no quadro 8 percebe-se a importância da utilização de termos menos comuns, pois o diagnóstico de interesse é retornado na segunda posição da lista representando uma boa recomendação, em relação aos termos da consulta.

4.2.7 Recomendações para Urticária

Para o sétimo teste com os sintomas da *Urticária* representando a queixa do paciente durante o atendimento, utilizou-se os seguintes termos: "*Superfície na pele vermelhos, coçam intensamente na pele*". O quadro 10 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.99460451967114	Constipação | resfriado
0.99163260272058	Galactosemia
0.98891887476699	Hipocalcemia
0.98859680941397	Urticária
0.98333261854842	Vaginismo
0.98023004729386	Hipoparatireoidismo
0.97979657434665	Queloide
0.9797869977507	Difteria
0.97963423274731	Melanoma
0.97715666295079	Gengivite

Quadro 10 - Teste de recomendação com sintomas da Urticária.

Fonte: Próprio Autor.

Um outro caso onde a recomendação apresentou um bom resultado devido ao uso de termos com forte valor semântico.

4.2.8 Recomendações para Febre Tifoide

Para o oitavo teste com os sintomas da *Febre tifoide* representando a queixa do paciente durante o atendimento, utilizou-se os seguintes termos: "*Febre alta, dor de cabeça, mal-estar, falta de apetite, retardamento do ritmo cardíaco, baço inchado, manchas rosadas, prisão de ventre, tosse seca*". O quadro 11 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.92296464046221	Febre tifoide
0.88533977778496	Mononucleose
0.88442825872615	DMRI
0.87760148754315	Pneumonia
0.87663529992	Varicocele
0.87006524476373	Clamídia
0.86611224902656	Vitiligo
0.85981021315512	Leucemia
0.85492444401414	Hipoparatireoidismo
0.854130025399	Hemocromatose

Quadro 11 - Teste de recomendação com sintomas da Febre Tifoide.
Fonte: Próprio Autor.

4.2.9 Recomendações para Pneumonia

Para o nono teste com os sintomas da *Pneumonia* representando a queixa do paciente durante o atendimento, utilizou-se os seguintes termos: *"Febre alta, Tosse, Dor no tórax, Alterações da pressão arterial, Confusão mental, Mal-estar, Falta de ar, muco purulento de cor amarelada ou esverdeada, Toxemia"*. O quadro 12 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.94967062110313	Leucemia

0.94910162437288	Pneumonia
0.93377778322505	Febre tifoide
0.93023526123248	Varicocele
0.93009956461434	Hemocromatose
0.92981783676831	Clamídia
0.92373621749417	Ansiedade
0.91865676366835	DMRI
0.91352715936492	Gripe H1N1
0.91197832874115	Anemia falciforme

Quadro 12 - Teste de recomendação com sintomas da Pneumonia.

Fonte: Próprio Autor.

4.2.10 Recomendações para Reumatismo

Para o decimo teste com os sintomas da *Pneumonia* representando a queixa do paciente durante o atendimento, utilizou-se os seguintes termos: *"Dores nas articulações, vermelhidão, calor, inchaço nas articulações, dores ao esticar os braços sobre a cabeça, dores ao elevar os ombros até tocar o pescoço"*. O quadro 13 apresenta a lista de resultados representando os possíveis diagnósticos.

Similaridade	Diagnósticos
0.95893700663175	Varicocele
0.95723486260981	Leucemia
0.9550897989195	Botulismo
0.95497290718909	Mononucleose
0.95170904179822	Vitiligo
0.94847701560753	Zika vírus
0.94615519286744	Poliomielite
0.94017475938303	Hipocondria
0.93809615940123	Hemocromatose
0.93565136575532	Hipocalemia

Quadro 13 - Teste de recomendação com sintomas da Reumatismo.

Fonte: Próprio Autor.

4.3 Resultados

A base de dados representando os prontuários de atendimentos dos pacientes foi constituída de registros coletados em sites da internet que continham doenças e seus respectivos sintomas. O universo da pesquisa foi representado por uma base de dados contendo 100 doenças com seus respectivos sintomas. De acordo com (GUIMARÃES, 2012; PROVDANOV; FREITAS, 2013; SILVA; KARKOTLI, 2011) universo da pesquisa representa o conjunto de seres animados ou não que apresenta pelo menos uma característica comum, sendo N o número total de elementos ou população.

Desse universo, serão utilizados 10 conjuntos como amostra representando os novos atendimentos, onde a doença representa o diagnóstico e os sintomas a queixa do paciente. Os conjuntos foram selecionados de aleatoriamente das ocorrências mais frequentes da atenção básica de saúde nos hospitais de acordo com (PIMENTEL et al., 2012). Quanto a amostra, (PROVDANOV; FREITAS, 2013) explica que é uma pequena parte do universo. A figura 22 apresenta um gráfico com a qualidade das recomendações dos testes de validação.

Figura 22 - Qualidade das recomendações.
Fonte: Próprio autor.

O total de termos presentes na base de dados foi 6344 em um total de 100 documentos, a amostra utilizada foi de 10% do universo estabelecido, a execução do algoritmo apresentou 7 resultados satisfatórios, ou seja, em 70% dos testes, retornou o diagnóstico de interesse entre as 10 melhores posições na lista de recomendações. No entanto nos outros 30% das execuções onde o resultado não foi satisfatório, observou-se que diagnóstico como o da gripe não foram bem classificados possivelmente por conterem nos sintomas que nesta pesquisa representaram as queixas do paciente,

continham termos muito comuns a todos os documentos da coleção, como por exemplo “febre”, “dor”, “cabeça” entre outros tiveram um grande número de repetições nos documentos da coleção.

Este é um comportamento esperado, visto que pela parte IDF do algoritmo, atua reduzindo a importância do termo que se repete em muitos documentos da coleção (BAEZA-YATES; RIBEIRO-NETO, 2013; KIDO; JUNIOR; MORIGUCHI, 2014; MARTHA; DE CAMPOS; SIGULEM, 2010). Diante de tal fato, presume-se que não é garantida a primeira posição no ranking de resultados, mesmo que se informe todos os termos da descrição de um diagnóstico. O que garante um bom posicionamento é sem dúvidas a utilização de termos menos genéricos para descrição dos itens da queixa dos pacientes.

Apesar de a base de testes conter apenas 100 registros, a especialização desta tende a melhorar a qualidade das recomendações, contudo o tamanho do documento também favorece melhores resultados, ou seja, quanto maior o documento, maior a chance de conter termos menos comuns e consequentemente melhorar seu posicionamento na lista de resultados das recomendações. Sendo o rendimento inicial do algoritmo de recomendações de 70% de acertos nas recomendações e em 28,57% dos casos de testes apresentou na primeira lista o diagnóstico de interesse, percebe-se ser uma estratégia eficiente na recomendação de diagnósticos.

5. CONSIDERAÇÕES FINAIS

O modelo de sistema proposto nesta pesquisa utilizou conceitos inerentes a recuperação de informação afim de classificar e recomendar diagnósticos pretéritos elaborados por outros médicos utilizando técnicas filtragem baseada em conteúdo, ou seja, o sistema foi capaz de comparar textos informados durante o atendimento de um novo paciente com os textos informados pelos médicos em seus atendimentos no passado. Esta estratégia apresentou uma ferramenta muito eficiente no que diz respeito a busca de informações em bases de dados existentes e de forma muito simples. Nesta pesquisa apresentou-se apenas o diagnostico como informação recuperada, contudo a sua aplicação pode ser ainda maior, no sentido de apresentar outras condutas médicas relacionadas aos registros recomendados.

Para alcançar o objetivo geral proposto neste trabalho foi desenvolvido um protótipo de prontuário eletrônico de pacientes em escala reduzida, que pode ser conferido no Apêndice C. Este por sua fez auxiliou na construção da base de dados que representou os atendimentos médicos dos pacientes, que mais tarde passaram pelo processo de indexação através da aplicação do algoritmo de geração de índices (pesos) TF-IDF implementado nesta pesquisa.

A similaridade dos conteúdos referente aos prontuários foi determinada através da formula dos cossenos, onde o peso dos termos da consulta que nesta pesquisa representou a queixa do paciente, foi submetida ao algoritmo e valores próximo de 1 (um) representou bons valores de similaridade e consequentemente, foram classificados de acordo com esta medida.

O universo da pesquisa foi de 100 prontuário cadastrados previamente na base de dados de testes, e apenas 10% deste universo constituiu a amostra para testar e validar a qualidade das recomendações do algoritmo. Nos testes foram encontrados 70% de aproveitamento das recomendações e apenas 30% não retornaram uma lista aceitável de opções. Muito provavelmente estes diagnósticos não encontrados apresentam muita semelhança semântica entre si e com todos os outros prontuários da coleção. Contudo com o crescimento da base de dados espera-se melhorar sensivelmente a qualidade das recomendações de acordo com o crescimento da base de dados (GARCIA; FROZZA, 2013).

O algoritmo implementado não tratou erros de digitação, sinônimos e/ou jargões profissionais utilizados na medicina. Estes valores dentro de um documento podem influenciar negativamente os resultados da busca durante o atendimento, contudo o sistema apresentou resultados satisfatórios.

Pode-se concluir que a recomendação de diagnóstico através da similaridade de conteúdo é uma técnica aceitável mesmo com poucas informações no banco de dados e mostrou-se bastante prática na pesquisa e classificação de resultados, podendo ser aplicada para apresentar outros procedimentos da conduta médica durante o seu atendimento.

6. REFERÊNCIAS

BAEZA-YATES; RIBEIRO-NETO. **Recuperação de Informação: conceitos e tecnologias das máquinas de busca**. [s.l: s.n.].

BAX, M. P. DESIGN SCIENCE: FILOSOFIA DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO E TECNOLOGIA. In: XV ENANCIB "ALÉM DAS NUVENS: EXPANDINDO AS FRONTEIRAS DA CIÊNCIA DA INFORMAÇÃO", 2014, Belo Horizonte. XV ENANCIB. Encontro Nacional de Pesquisa em Ciência da In. p. 3883–3903, 2014.

BRITO, A. G. PROPOSTA DE MODELO DE RECOMENDAÇÃO DE CONTEUDO BASEADO EM ARQUIVOS DE LEGENDAS DE FILMES E SERIES. 2016.

CAPPELLA, J. N.; YANG, S.; LEE, S. Constructing Recommendation Systems for Effective Health Messages Using Content, Collaborative, and Hybrid Algorithms. **The ANNALS of the American Academy of Political and Social Science**, v. 659, n. 1, p. 290–306, 2015.

CAZELLA, S. C.; NUNES, M.; REATEGUI, E. A Ciência da Opinião: Estado da arte em Sistemas de Recomendação. **CSBC XXX Congresso da SBC Jornada de Atualização de InformáticaJAI**, p. 161–216, 2010.

COSTA, E.; AGUIAR, J.; MAGALHÃES, J. Sistemas de Recomendação de Recursos Educacionais: conceitos, técnicas e aplicações. **Jaie 2013**, n. Cbie, p. 57–78, 2013.

COSTA, I. DE O. MODELOS PARA ANÁLISE DE DISPONIBILIDADE EM UMA PLATAFORMA DE MOBILE BACKEND AS A SERVICE. 2015.

FADAEE, S. S.; SOUFIANI, H. A.; SUNDARAM, R. Chiron: A Robust Recommendation System. 2016.

FERNEDA, E. Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003.

GARCIA, C. A; FROZZA, R. Sistema de Recomendação de Produtos Utilizando Mineração de Dados. **17**, p. 78–90, 2013.

GUIMARÃES, P. R. B. Métodos Quantitativos Estatísticos. p. 252, 2012.

IRACICABA, P. Recuperação de documentos texto usando modelo probabilistico estendido. 2006.

KIDO, G. S.; JUNIOR, S. B.; MORIGUCHI, S. N. Comparação entre TF-IDF e LSI para pesagem de termos em micro-blog. n. May, 2014.

MAIA, L. C. G.; SOUZA, R. R. Medidas de similaridade em documentos eletrônicos 1. **IX ENANCIB: Diversidade Cultural e Políticas de Informação**, v. 1, n. 1, p. 1–15, 2013.

MARIN, H. D. F. Sistemas de informação em saúde: considerações gerais. **Journal of Health Informatics**, v. 2, n. 1, p. 20–24, 2010.

MARTHA, A. S.; DE CAMPOS, C. J. R.; SIGULEM, D. Recuperação de Informações em Campos de Texto Livres de Prontuários Eletrônicos do Paciente Baseada em Semelhança Semântica e Ortográfica. v. 2, n. 3, p. 63–71, 2010.

MIGUEL, A.; COSTA, G. AToMRS : Uma Ferramenta para Monitorizar o Desempenho de Sistemas de Recomendação. 2016.

PAIK, J. H. A Novel TF-IDF Weighting Scheme for Effective Ranking. **Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval**, p. 343–352, 2013.

PEREZ, G.; ZWICKER, R. Fatores determinantes da adoção de sistemas de informação na área de saúde: um estudo sobre o prontuário médico eletrônico. **RAM. Revista de Administração Mackenzie (Online)**, v. 11, n. 1, p. 174–200, 2010.

PETERS, A. C. et al. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. **Anais do XII Congresso Brasileiro de Informática em Saúde**, n. 1, 2010.

PIMENTEL, Í. R. S. et al. Caracterização da demanda em uma Unidade de Saúde da Família. **Revista Brasileira de Medicina de Família e Comunidade**, v. 6, n. 20, p. 175–181, 2012.

PROVDANOV, C. C.; FREITAS, E. C. DE. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. [s.l: s.n.].

RAMOS J. G. A. Algoritmos Colaborativos para Sistemas de Recomendação. 2010.

SALVADORI, I. L. DESENVOLVIMENTO DE WEB APIS RESTFUL SEMÂNTICAS BASEADAS EM JSONo. **Igarss 2014**, n. 1, p. 1–5, 2014.

SILVA, R.; KARKOTLI, G. Manual De Metodologia Científica Do Usj - 2011 - 1. 2011.

SOUSA, O. DE; TABOSA, H. R. A EFICACIA DOS MODELOS DE RECUPERAÇÃO DE INFORMAÇÕES: UM ESTUDO PARTICULARIZADO NA COMUNICAÇÃO CIENTIFICA NA WEB. 2015.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161–173, 2006.

U, R. L. A Survey on Recommender System. v. 1, n. 6, p. 50–57, 2013.

VERBERT K.; MANOUSELIS, N. . O. X. . W. M. . D. H. . B. I. . D. E. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. **Learning Technologies, IEEE Transactions on**, v. 5, n. 4, p. 318–335, 2012.

VIEIRA, F.; NUNES, M. DICA: Sistema de Recomendação de Objetos de Aprendizagem Baseado em Conteúdo. **Scientia Plena**, v. 8, p. 1–10, 2012.

WEITZEL, L.; PALAZZO, J.; OLIVEIRA, M. DE. Sistemas de recomendação de informação em saúde baseado no perfil do usuário. **Journal of Health Informatics**, p. 1–7, 2010.

YANG, S. et al. Application of Statistical Relational Learning to Hybrid Recommendation Systems. 2016.

Apêndice A - Criando o projeto no Firebase

Para a criação do projeto no Firebase fez-se necessário a criação de uma conta no serviço, utilizou-se o plano gratuito por oferecer boas configurações. Após a criação da conta e escolha do plano foi apresentada a seguinte interface para a correta configuração da aplicação a ser criada.

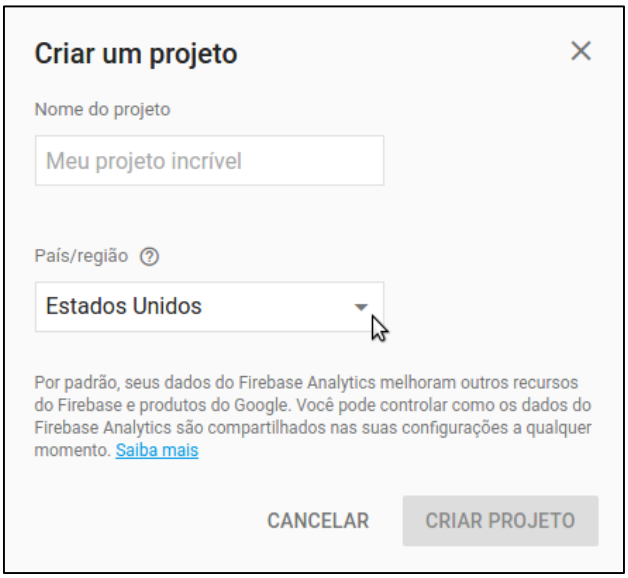

Figura 23 - Definição do projeto no Firebase

Fonte: Próprio autor.

Após a criação do projeto no *Firebase* foi apresentada a seguinte interface mostrando as ferramentas disponíveis para o referido projeto. A figura 23 apresenta a interface do console de configuração do Firebase.

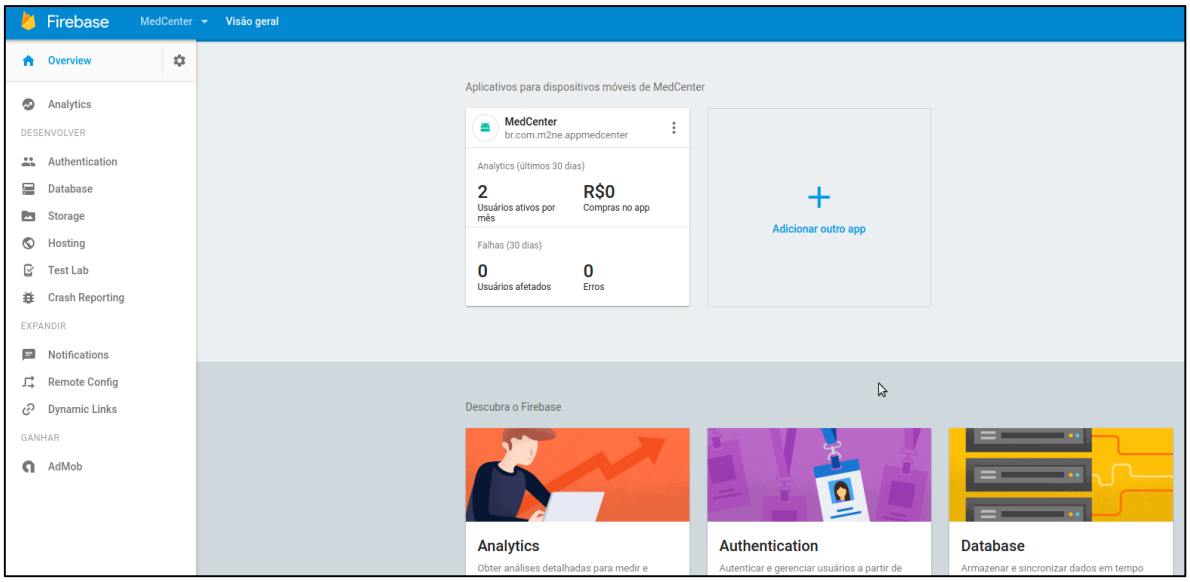

Figura 24 - Ferramentas do Firebase

Fonte: Próprio autor.

Apenas as ferramentas *Authentication* e *Database* foram utilizadas nesta pesquisa.

- *Authentication* – Interface para autenticação de usuários.
- *Database* – Interface para persistência de dados.

Para a conclusão do webservice foi preciso fazer a integração por meio do SDK do Firebase com o projeto do protótipo de prontuário eletrônico para que ambos possam trocar informações. A figura 24 apresenta as dependências necessárias para o correto funcionamento do Firebase.

```
dependencies {
    compile fileTree(include: ['*.jar'], dir: 'libs')
    androidTestCompile('com.android.support.test.espresso:espresso-core:2.2.2',
        exclude group: 'com.android.support', module: 'support-annotations'
    })
    compile 'com.android.support:appcompat-v7:24.2.1'
    compile 'com.android.support:design:24.2.1'
    compile 'com.google.firebase:firebase-auth:9.6.1'
    compile 'com.google.firebase:firebase-database:9.6.1'
    compile 'com.google.firebase:firebase-storage:9.6.1'
    compile 'com.android.support:support-v4:24.2.1'
    compile 'com.google.android.gms:play-services:9.6.1'
    compile 'com.google.firebase:firebase-crash:9.0.2'
    compile 'com.google.firebase:firebase-messaging:9.0.2'
    compile 'com.firebaseui:firebase-ui:0.4.0'
    compile 'com.facebook.android:facebook-android-sdk:[4,5)'

    testCompile 'junit:junit:4.12'
    compile 'com.google.code.gson:gson:2.8.0'
}
```

Figura 25 - Dependências do Firebase

Fonte: Próprio autor.

Apêndice B – Requisitos e modelagem do protótipo

Na construção do software do prontuário eletrônico de paciente foi elaborado uma lista de requisitos de funcionalidade contempladas pelo software conforme ilustra o quadro 14 abaixo:

RF001	Registrar Anamnese do paciente
RF002	Registrar Cirurgias do paciente
RF003	Registrar Medicamentos do paciente
RF004	Registrar Diagnósticos do paciente
RF005	Registrar Internações do paciente
RF006	Registrar dados psicológicos do paciente
RF007	Registrar dados do atendimento social
RF008	Registrar dados da evolução diária

Quadro 14 - Lista de requisitos funcionais modulo prontuário eletrônico do paciente

Fonte: Próprio autor.

Doravante, para melhor compreender a função do médico e outros membros da equipe durante a utilização do prontuário eletrônico foi elaborado um diagrama da Linguagem de Modelagem Unificada (UML) denominado caso de uso, conforme figura 25.

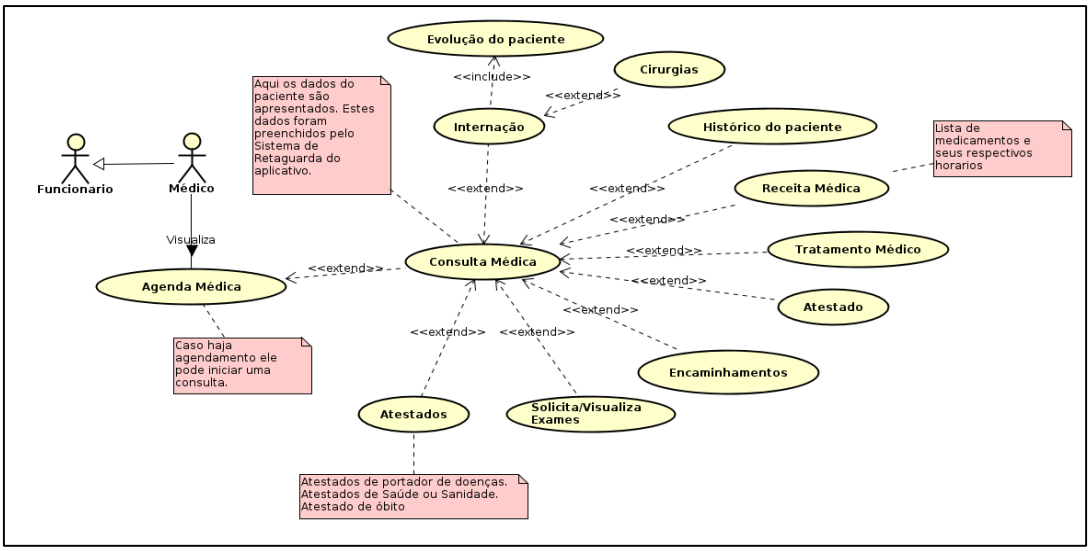

Figura 26 - Diagrama de caso de uso prontuário eletrônico do paciente

Fonte: Próprio autor.

Outro diagrama da UML bastante comum e muito utilizado para representar a colaboração entre os vários objetos que farão parte do sistema é o diagrama de classes. A figura 26 apresenta o diagrama de classes do modulo prontuário eletrônico.

Fonte: Próprio autor.

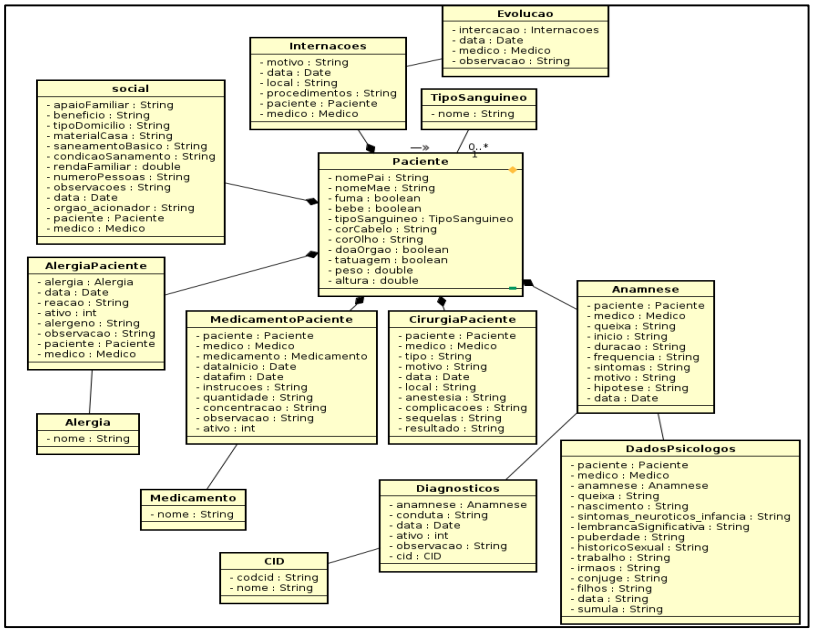

Figura 27 - Diagrama de classes prontuário eletrônico do paciente.

Fonte: Próprio autor.

No entanto para o correto funcionamento do software prontuário eletrônico de paciente, faz-se necessário o desenvolvimento de mais dois módulos auxiliares denominados modulo administrador e modulo de atendimento.

O modulo administrador é responsável por fornecer interfaces de configuração do sistema proposto, bem como ferramentas de cadastro de informações que auxiliarão no funcionamento dos demais módulos. O quadro 15 apresenta os requisitos funcionais do modulo administrador.

Requisito	**Descrição**
RF009	Cadastro de Usuários.
RF010	Cadastro de grupos de usuários.
RF011	Cadastro de Especialidade Médica
RF012	Cadastro de Cidades
RF013	Cadastro doenças de acordo com a CID 10
RF014	Cadastro de Raça
RF015	Cadastro de Cor
RF016	Cadastro de Estado civil
RF017	Cadastro de Médicos
RF018	Cadastro de Escolaridade
RF019	Cadastrar medicamentos
RF020	Cadastrar Alergias

Quadro 15 - Requisitos funcionais do modulo administrador

Fonte: Próprio autor.

As figuras 27 e 28 apresentam os diagramas de casos de uso e diagrama de classes do modulo administrador.

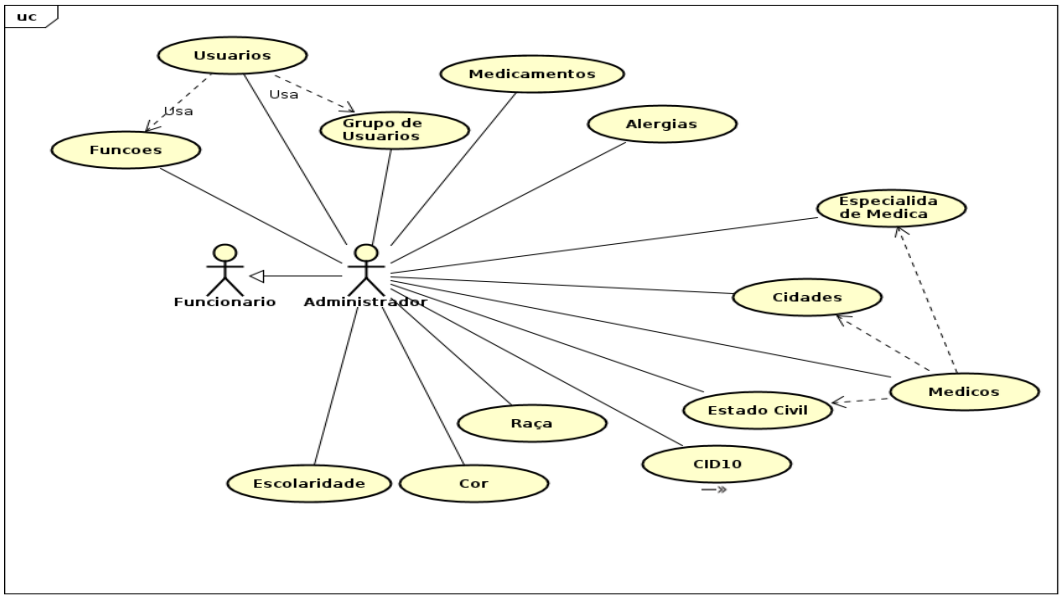

Figura 28 - Diagrama de caso de uso modulo administrador

Fonte: Próprio autor.

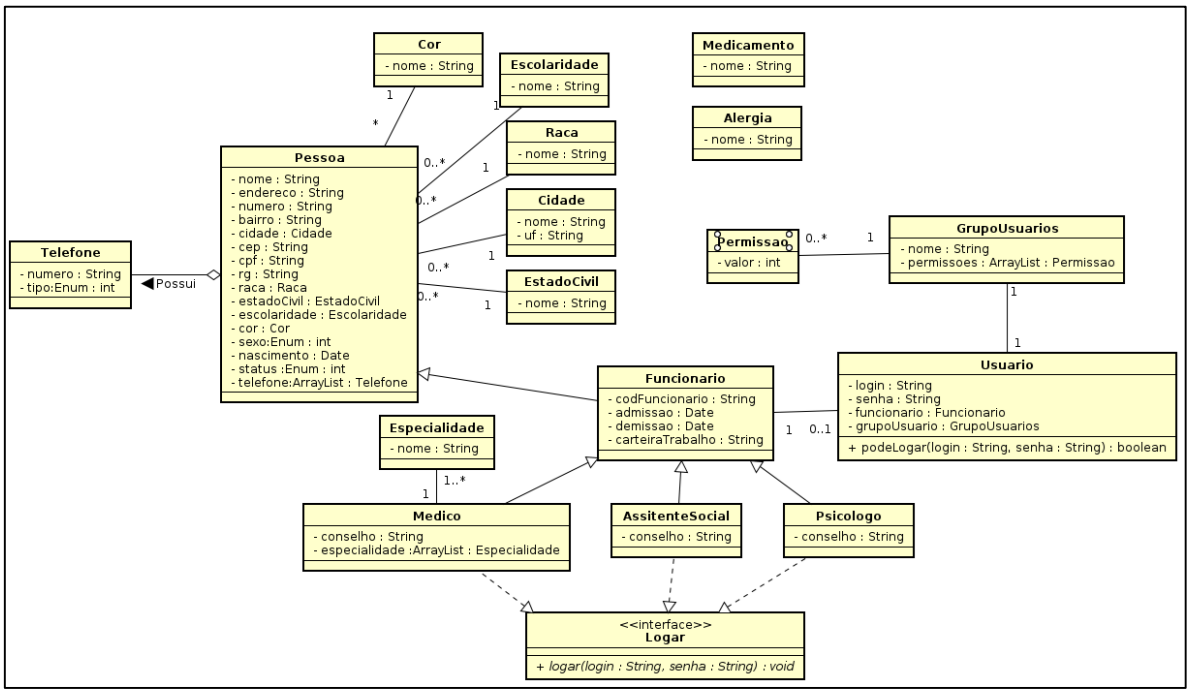

Figura 29 - Diagrama de classes modulo administrador

Fonte: Próprio autor.

O modulo de atendimento é responsável pelas interfaces para cadastramento de pacientes e organização da fila de atendimento através do agendamento de consultas para um médico, conforme lista de requisitos do quadro 16.

RF021	Cadastrar e visualizar de Pacientes
RF022	Agendar e visualizar Consulta de pacientes
RF023	Registrar e visualizar Resultado de Exames

Quadro 16 - Requisitos funcionais modulo de atendimento de pacientes.

Fonte: Próprio autor.

As figuras 29 e 30 apresentam os diagramas de caso de uso e diagrama de classes do modulo de atendimento de paciente.

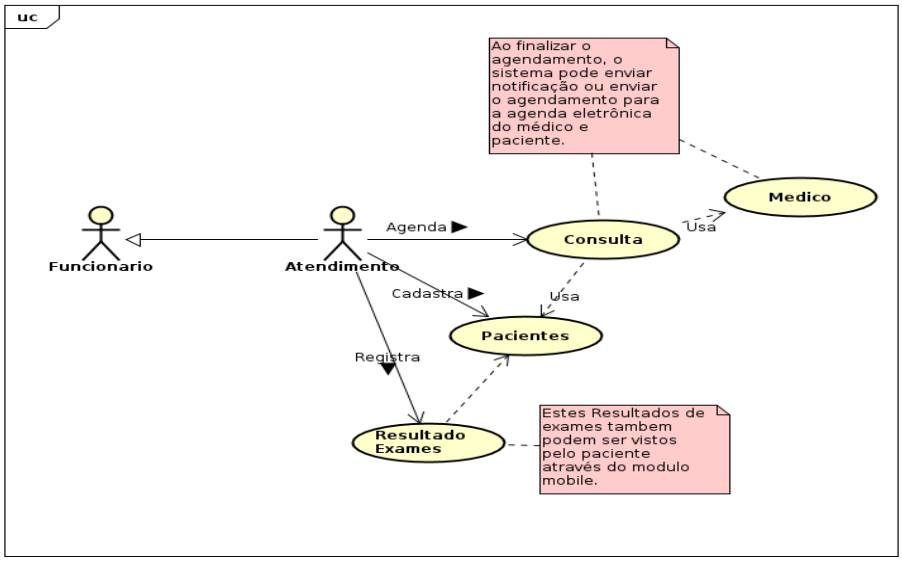

Figura 30 - Diagrama de casos de uso modulo de atendimento

Fonte: Próprio autor.

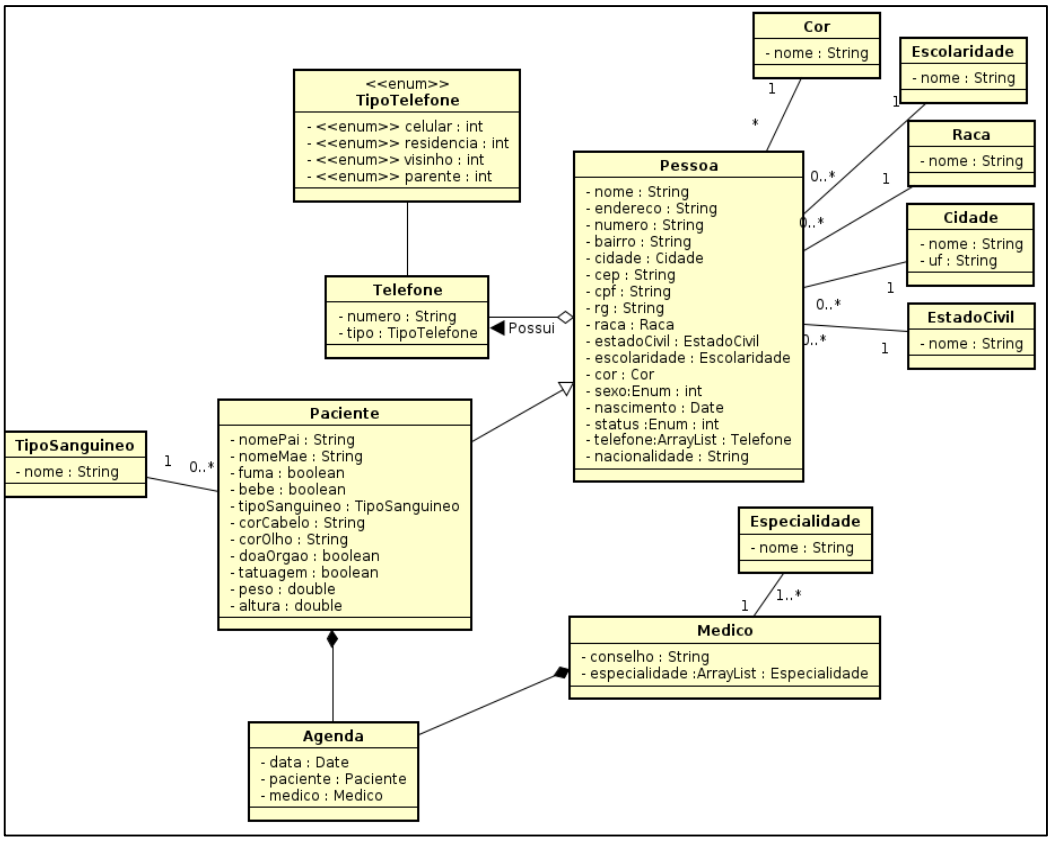

Figura 31 - Diagrama de classes modulo de atendimento

Fonte: Próprio autor.

Apêndice C - Requisitos do protótipo do prontuário eletrônico

Para a construção do protótipo do aplicativo de prontuário eletrônico utilizou-se técnicas de engenharia de software reversa através de software já existentes no mercado para determinar os requisitos (em sua forma reduzida) a serem implementados. Abaixo encontra-se os quadros com os respectivos requisitos funcionais contemplados no protótipo.

Identificação do paciente	Informações necessárias
Requisito utilizado para inserção das informações de cadastro do paciente na base de dados do sistema.	Nome E-mail Nascimento Sexo Observação Telefone fixo Telefone celular

Quadro 17 – Identificação do paciente

Fonte: Próprio autor.

Identificação do médico	Informações necessárias
Requisito utilizado para inserção das informações de cadastro do médico na base de dados do sistema.	Nome E-mail Tratamento Sexo Conselho Registro Telefone celular

Quadro 18- Identificação do médico

Fonte: Próprio autor.

Login de acesso ao sistema	Informações necessárias

Requisito utilizado acessar o sistema de prontuário eletrônico.	Nome E-mail

Quadro 19 - Login de acesso ao sistema

Fonte: Próprio autor.

Lista de pacientes	Informações necessárias
Requisito utilizado para listar em ordem alfabética os pacientes cadastrados no sistema. Obs.: Nesta lista pode-se selecionar o paciente para início de um novo prontuário.	

Quadro 20 - Lista de pacientes

Fonte: Próprio autor.

Novo prontuário	Informações necessárias
Requisito utilizado para criação de um novo prontuário eletrônico do paciente. Neste requisito lista-se as seções disponíveis para o médico.	Anamnese Exame Físico Sinais Vitais Hipótese diagnostico Prescrição Evolução Atestado

Quadro 21 - Novo prontuário

Fonte: Próprio autor.

Ficha Anamnese	Informações necessárias

Requisito utilizado para o cadastro dos dados da ficha anamnese do paciente.	Queixa principal História do paciente Problemas renais Problemas articulares ou reumatismo Problemas cardíacos Problemas gastrointestinais Problemas alérgicos Se é portador de hepatite Gravidez Se é diabético Problemas com cicatrização Relato de uso de medicamentos

Quadro 22 - Ficha anamnese
Fonte: Próprio autor.

Apenas os dados da ficha anamnese foram utilizados para esta pesquisa. As demais seções foram implementadas, mas por se tratarem basicamente de medidas matemáticas possuem pouca relevância para as futuras recomendações.

Implementação dos requisitos

Após a definição dos requisitos do protótipo, iniciou-se a etapa de implementação que resultou na construção das classes que contém as funcionalidades e informações apresentadas em cada requisito contemplado. Nesta etapa optou-se pela utilização de um paradigma de programação orientado a objetos para construção das classes entidades e estratégia de persistência dos dados. A linguagem de programação utilizada foi o Java na sua versão 7. A figura 31 apresenta a classe entidade referente a identificação médico.

```
public class Medico {
    private String uid;
    private String nome;
    private String email;
    private String celular;
    private String registro;
    private String tratamento;
    private String conselho;
    private String sexo;

    public Medico() {

    }
```

Figura 32 - Classe entidade identificação do médico.

Fonte: Próprio autor.

A figura 32 e 33 apresenta a classe entidade referente aos requisitos de identificação do paciente e ficha anamnese.

```
public class Pacientes implements Salvar {

    private String uid;
    private String nome;
    private String nascimento;
    private String email;
    private String telefone;
    private String celular;
    private String observacao;
    private String sexo;

    public Pacientes() {

    }
```

Figura 33 - Classe entidade identificação do paciente

Fonte: Próprio autor.

```
public class Anamnese {
    private String queixa;
    private String historico;
    private String problemasRenais;
    private String problemasArticulares;
    private String problemaCardiaco;
    private String problemaRespiratorio;
    private String problemaGastrico;
    private String alergia;
    private boolean hepatite;
    private boolean diabete;
    private boolean gravidez;
    private boolean problemaCicatrizacao;
    private String medicamentos;
    private Object timestamp;
```

Figura 34 - Ficha anamnese

Fonte: Próprio autor.

As classes entidades foram utilizadas no projeto para representar os dados definidos em cada requisito contemplado e que serão preenchidos pelos usuários através das interfaces que serão descritas mais adiante. No entanto para que estes dados sejam gravados permanentemente é preciso a criação de uma camada de persistência. Para esta camada utilizou-se o (*Software Development Kit*) SDK Firebase, um serviço de (*Mobile Backend as a service*) MBAAS desenvolvido e mantido pela Google, este recurso apresenta uma significativa redução nos custos de implementação por fornecer um conjunto de API´s e SDK´s de forma gratuita e por tempo indeterminado.

O método **salvar** presente em algumas classes entidades se encarrega da persistência das informações no banco de dados do sistema, esta faz chamadas aos métodos existentes em uma classe denominada Firebase. A classe entidade Pacientes implementa este método conforme a figura 34.

```
public boolean salvar(){
    DatabaseReference db = Firebase.getRefs(Constantes.FIREBASE_CHILD_PACIENTE);
    try {
        this.uid = db.push().getKey();
        db.child(this.uid).setValue(this);
        return true;
    }
    catch(Exception e){
        FirebaseCrash.log(e.getMessage());
        return  false;
    }
}
```

Figura 35 - Estratégia de persistência de dados classe entidade pacientes.
Fonte: Próprio autor.

A classe Firebase contém as chamadas das bibliotecas do SDK do Firebase, esta classe se encarrega de sincronizar os dados informados nas classes entidades com o modulo de webservice que será descrito na próxima seção. A figura 35 apresenta sua implementação.

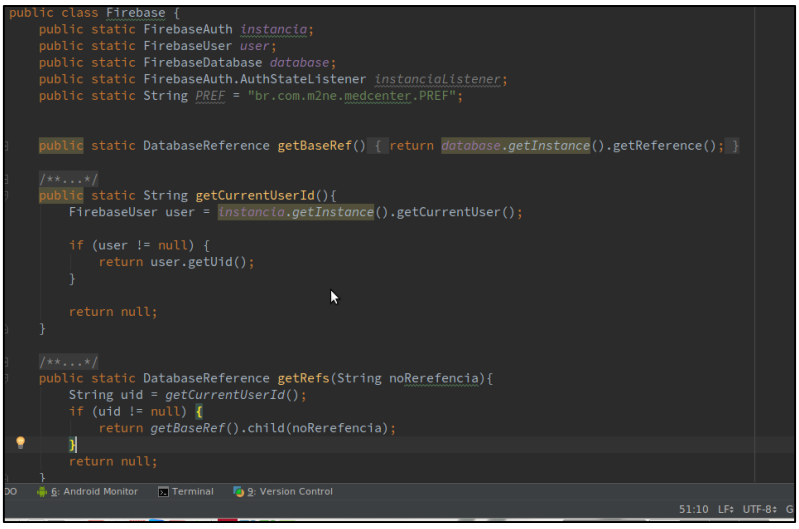

Figura 36- Implementação da classe Firebase.
Fonte: Próprio autor.

Implementação das interfaces do protótipo do aplicativo.

A interface do protótipo foi construída utilizando recursos nativos da ferramenta *Android Studio.* A primeira interface implementada foi o menu principal do protótipo cuja finalidade é dar suporte ao usuário na escolha das ferramentas a serem usadas. A figura 36 apresenta esta interface.

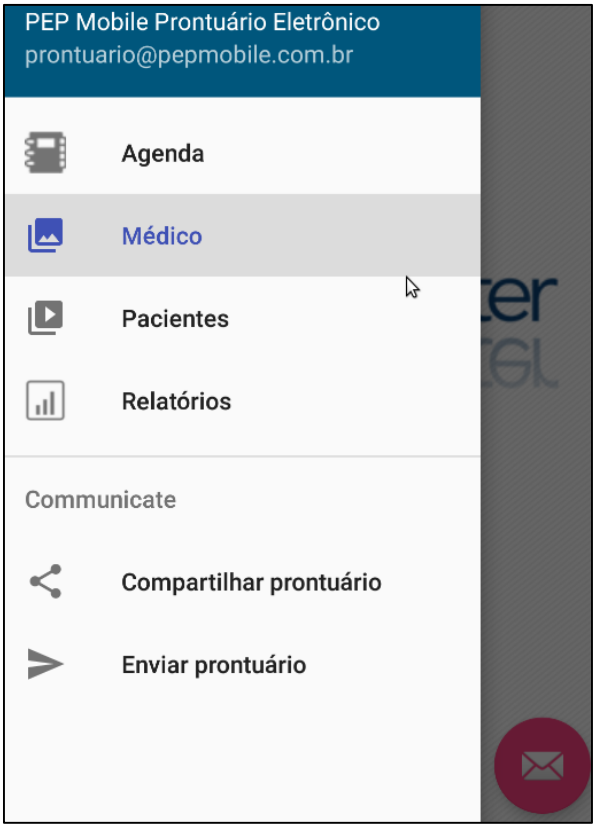

Figura 37 - Menu do protótipo de prontuário eletrônico.

Fonte: Próprio autor.

As figuras 37 e 38 apresentam a interface de identificação do médico e paciente respectivamente.

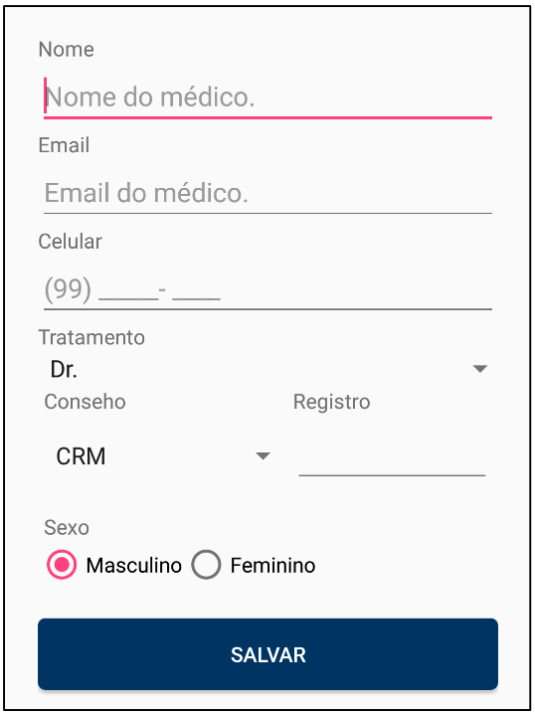

Figura 38 - Identificação do médico

Fonte: Próprio autor.

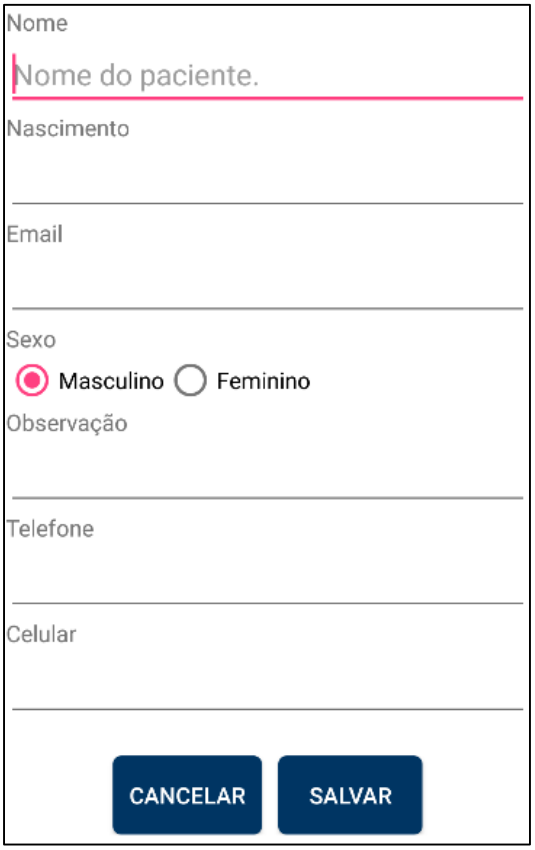

Figura 39 - Identificação do paciente

Fonte: Próprio Autor.

A figura 39 apresenta a lista de paciente cadastrados no sistema.

Figura 40 - Lista de pacientes cadastrados.

Fonte: Próprio Autor.

A figura 40 apresenta as opções disponíveis para criação de um novo prontuário eletrônico do paciente.

Figura 41 - Opções do prontuário eletrônico

Fonte: Próprio autor.

A figura 41 apresenta parte superior da interface de preenchimento da ficha anamnese.

Figura 42 - Parte superior da ficha anamnese

Fonte: Próprio autor.