Explorando Intenções na Recuperação de Informação em Prontuários Eletrônicos de Pacientes

Edemar Mendes Perciani

Dezembro / 2016

Dissertação de Mestrado em Ciência da Computação

Explorando Intenções na Recuperação de Informação em Prontuários Eletrônicos de Pacientes

Esse documento corresponde a Dissertação apresentada à Banca Examinadora para a defesa de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista.

Campo Limpo Paulista, 02 de dezembro de 2016.

Edemar Mendes Perciani

Rodrigo Bonacin (Orientador)
Julio Cesar dos Reis (Co-orientador)

Dados Internacionais de Catalogação na Publicação (CIP)

Câmara Brasileira do Livro, São Paulo, Brasil

Perciani, Edemar Mendes

Explorando intenções na recuperação de informação em prontuários eletrônicos de pacientes / Edemar Mendes Perciani. Campo Limpo Paulista, SP: FACCAMP, 2016.

Orientador: Prof. Dr. Rodrigo Bonacin

Co-orientador: Prof. Dr. Julio Cesar dos Reis

Dissertação (mestrado) – Faculdade Campo Limpo Paulista – FACCAMP .

1. Recuperação de informação. 2. Prontuário eletrônico do paciente. 3. Expansão de consultas. 4. Intenção. 5. Ilocução. 5. Semiótica organizacional. 6. Teoria dos atos da fala. I. Bonacin, Rodrigo. II. Reis, Julio Cesar dos. III. Faculdade Campo Limpo Paulista. VI. Título.

CDD-005.75

Dedicatória

A minha Mãe que esteve sempre ao meu lado, proporcionando-me além de extenso carinho e amor, os conhecimentos da integridade, da perseverança e de procurar sempre em Deus à força maior para o meu desenvolvimento como ser humano. Por essa razão, gostaria de dedicar e reconhecer a você, minha imensa gratidão amor.

À Deus, dedico o meu agradecimento maior, porque atribuo tudo o que tenho na minha vida a ele. Nas minhas orações, uma a uma, todas são sempre ouvidas e ao seu tempo são sempre respondidas, em especial a de ser Mestre.

Agradecimentos

Aos Orientadores Dr. Rodrigo Bonacin e Professor Dr. Julio Cesar dos Reis que se empenharam de forma significativa com conhecimentos e ideias compondo grande parte do crédito do trabalho realizado. O resultado maior desta orientação se resume em uma amizade que vai além dos portões da FACCAMP.

A todos aqueles que fizeram parte da minha vida dentro e fora do mestrado. Uma pesquisa científica não se faz da noite para o dia, tão pouco se resume a um conhecimento isolado. Desta forma, aprendi muito com meus amigos que estiveram junto comigo ao longo desta jornada. Eu gostaria de poder citar cada um, entretanto a lista seria longa demais se eu fosse contar o que cada um de vocês significa para mim. Então deixo minha eterna gratidão e amizade.

A todos os professores do Programa de Mestrado em Ciência da Computação da FACCAMP, em especial a Prof. Dr. Ana Maria Monteiro, a Prof. Dr. Maria do Carmo Nicolette ao Prof. Dr. Osvaldo Luis de Oliveira.

“Combati o bom combate, terminei minha carreira, guardei a fé”.

Timóteo 4:7

Resumo: Prontuários Eletrônicos de Pacientes informatizam dados de pacientes para facilitar o acesso e aprimorar o desenvolvimento de tratamentos de saúde. Apesar dos potenciais benefícios para os pacientes, profissionais da área ainda encontram dificuldades na seleção de documentos relevantes, em bases volumosas, para suas atividades de pesquisas, gestão e práticas clínicas. Técnicas de recuperação de informação desenvolvem um papel central em máquinas de busca. Embora métodos têm sido investigados para considerar o significado dos termos em consultas visando maximizar a relevância e abrangência dos documentos retornados, esta pesquisa identificou que poucos estudos exploram o conceito de intenção, como ações propositais dos usuários, na recuperação de informação. Esta dissertação objetiva investigar um mecanismo de recuperação de informação que explora a modelagem formal de dimensões relacionadas ás intenções em prontuários. Para esse fim, a pesquisa se fundamenta no referencial teórico da Semiótica Organizacional e Teoria dos Atos da Fala para estruturar categorias de intenções. Este trabalho enfrenta desafios para determinar meios de efetuar seleção e ordenação dos resultados de busca considerando explicitamente intenções declaradas pelos usuários. Investigamos alternativas para anotar significados e intenções em documentos médicos descritos em linguagem natural por meio da análise de termos recorrentes para expressar intenções no domínio. Com base nas anotações obtidas, definimos e implementamos um algoritmo para filtrar e determinar a ordem dos resultados de busca de acordo com tipos de intenção declarados como parâmetros na consulta do mecanismo de busca. Esta pesquisa alcançou o desenvolvimento de um sistema de recuperação de informação. Para avaliar a proposta, conduzimos um estudo usando uma base real de prontuários com 13.300 registros, em dois cenários de busca envolvendo profissionais da área da saúde. A análise dos resultados, por meio de medidas objetivas de precisão e cobertura, demonstra o potencial da solução quando aplicada a cenários complexos de busca.

Palavras-chave: Recuperação de Informação, Prontuário Eletrônico do Paciente, Expansão de Consultas, Intenção, Ilocução, Semiótica Organizacional, Teoria dos Atos da Fala

Abstract: *Electronic Health Records bring to digital format the patients' information in order to facilitate and improve the development of medical treatments. Despite the potential benefits to patients, health care professionals still face difficulties in the selection of relevant documents in huge repositories to aid their activities of research and managements as well as clinical practices. Information retrieval techniques play a key role in search engines. Although methods have been investigated to consider the meanings of terms informed by users to maximize the relevance and coverage of the retrieved documents, this research identified that few studies explore intentions, as explicit users' actions, in information recovery. This MSc. dissertation aims at investigating the development of an innovative information retrieval mechanism that explores the formal representation of intentions on Electronic Health Records. To this end, this research relies on Organizational Semiotics and Speech Acts Theory to categorize several types of intentions. This work faces issues to determine the selection and ranking of search results by taking users' declared intentions into account explicitly. We investigate alternatives for annotating the meanings and intentions in non-structured texts of medical records described in natural language by analyzing the recurrent terms to express intentions in the field. Based on the obtained annotations, we define an algorithm to filter and order search results based on intention classes declared as query parameters in the search mechanism. This research achieved the thoroughly development of an information recovery system. To evaluate the proposal, we conducted an experimental study using a real repository of Electronic Health Records containing 13.300 documents. Two search scenarios were defined involving health care professionals. Obtained results analyzed with measures of precision and recall demonstrate the efficiency of the solution when applied to complex search scenarios.*

Keywords: *Information Retrieval, Electronic Health Record, UMLS, Query expansion, Intentions, Illocutions, Organizational Semiotics, Speech Acts Theory*

Sumário

Glossário

CFM - *Conselho Federal de Medicina*

CID10 - *Código Internacional de Doenças*

CKM - *Clinical Knowledge Manager*

CUI - *Concept Unique Identifier*

DM2 - *Diabetes mellitus tipo 2*

EHR - *Electronic Health Records*

ESF - *Estratégia de Saúde da Família*

HL7 - *Health Level Seven International*

LOINC - *Logical Observation Identifiers Names and Codes*

MeSH - *Medical Subject Headings*

NLM - *National Library of Medicine*

OWL - *Ontology Web Language*

openEHR - *Open Eletronic Health Records*

PEP - *Prontuário Eletrônico do Paciente*

PLN - *Processamento de Linguagem Natural*

RDF - *Resource Description Framework*

SOC - *Sistema de Organização de Conhecimento*

TIC - *Tecnologia da Informação e Comunicação*

UMLS - *Unified Medical Language System*

UPA - *Unidade de Pronto Atendimento – Urgência e Emergência*

UTS - *UMLS Terminology Services*

W3C - *World Wide Web Consortium*

XML - *Extensible Markup Language*

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

No passado, as informações sobre o quadro clínico de pacientes eram armazenadas em prontuários físicos na forma de anotações em papel. O advento e evolução das Tecnologias de Informação e Comunicação (TICs) e sua aplicação na área de Informática Médica tem permitido que os prontuários médicos deixem de serem documentos físicos em arquivos de papéis. O Prontuário Eletrônico do Paciente (PEP[1]) torna digital as informações do paciente e possibilita o armazenamento em sistemas computacionais. Entretanto, o grande volume de documentos digitais atuais gera dificuldades para os profissionais de saúde analisarem o conjunto de dados disponível.

Esta situação requer mecanismos tecnológicos e algoritmos computacionais que permitam a recuperação adequada de documentos para identificar os mais relevantes a uma consulta do usuário. Avanços recentes na computação utilizam modelos subjacentes que representam a semântica em textos pouco estruturados descritos em linguagem natural para recuperar os documentos mais relevantes considerando a semântica dos termos da consulta (Hildebrand, Ossenbruggen & Hardman, 2007).

Essas propostas ainda apresentam restrições por não considerarem as intenções dos usuários ao produzir e recuperar informação. Detectar e representar intenções, e outros aspectos pragmáticos da linguagem humana, é um desafio atual para a área de Ciência da Computação. Isso é particularmente importante no contexto de PEPs na área da saúde, onde informações são expressas através de textos livres não estruturados descritos em linguagem natural. O processo de recuperação de informação pode se beneficiar do uso de intenções explicitamente declaradas (anotadas) no texto para a desambiguação de termos segundo um contexto (Bonacin *et al.*, 2013). Contudo, esse tema tem sido pouco estudado na literatura.

Esta dissertação visa investigar um método original de recuperação de informação através da representação e uso de intenções em algoritmos de busca. Mais

[1] Nesta dissertação é utilizado Prontuário Eletrônico do Paciente (PEP) como tradução do termo em inglês *Electronic Health Record* (EHR), amplamente utilizado na literatura de Informática Médica.

precisamente, este trabalho visa recuperar informação voltada para fins de pesquisa acadêmica, auditorias e gestão de saúde que busquem selecionar relatos de pacientes em PEPs na forma de texto livre. Buscas com este propósito visam explorar os procedimentos e exames descritos em PEPs que possuem diagnósticos relevantes para a confirmação de patologias, tratamentos, métodos alternativos de tratamentos entre outros. Para esse fim, objetiva-se explorar a Teoria dos Atos de Fala (Austin, 1962; Searle, 1976). Está fora do escopo deste trabalho, lidar com informações previamente codificadas e estruturadas em base de dados, tais como tabelas de procedimentos ambulatoriais e de internação. Também não se pretende focar em apoio ao atendimento clínico imediato com dados sobre um único paciente.

A principal contribuição desta dissertação é a concepção e desenvolvimento de um mecanismo de recuperação de informação que explora a modelagem computacional das intenções em PEPs. Implementamos a proposta em um sistema de *software* que é avaliado em estudo de caso na área médica com base em dados do sistema *Prontuário Digital* (Sistema Próprio) implantado em uma Unidade de Pronto Atendimento (UPA) no interior do Estado de São Paulo.

A avaliação explora cerca de 13.300 prontuários com o apoio de uma equipe médica multidisciplinar composta por um médico alergologista, três enfermeiros e um técnico de enfermagem. Os cenários analisados demostram melhorias nas medidas de precisão e cobertura de documentos obtidos nos resultados de buscas em relação aos métodos comumente utilizados para a recuperação de PEPs.

1.1. Problemática e Justificativa

Devido ao grande acumulo de informação em formato digital, mecanismos de recuperação de informação sobre PEPs desempenham um papel essencial no apoio à pesquisa por novos tratamentos e para o melhor atendimento ao paciente (Chen, Chung & Lin, 2012). A grande quantidade de documentos dificulta os usuários de julgarem e selecionarem os resultados relevantes dos pouco relevantes.

O problema da recuperação de informação envolve retornar uma lista ordenada de resultados com os documentos mais relevantes no topo da lista. Em uma busca o usuário utiliza, tradicionalmente, um conjunto de palavras-chave para expressar suas

necessidades. Para lidar com esse problema, diversos modelos matemáticos foram propostos para quantificar a relevância dos termos através de análises léxico-sintáticas e determinar os documentos mais relevantes para uma dada consulta. Usualmente, a avaliação das técnicas envolve quantificar através de métricas objetivas a qualidade dos resultados retornados. As medidas de avaliação mais exploradas são as de *Precisão, Cobertura* e *Medida-f.*

Pesquisas recentes na área de *Web* Semântica (*Dong*, Hussain, & Chang, 2008) têm estudado métodos para a construção de modelos e mecanismos de busca que considerem as representações computacionais sobre o significado da informação. Nesse sentido, modelos de representações de conhecimento, como o UMLS (*Unified Medical Language System*)[2] na área médica, são utilizados como artefatos computacionais subjacentes que contribuem na representação formal de conceitos para os sistemas. Entre os possíveis modelos de representações de conhecimento estão as ontologias, que são capazes de representar uma estrutura de relacionamentos semânticos de conceitos e podem superar deficiências dos métodos que apenas desenvolvem processamento léxico-sintático da informação (Cenan, 2008).

Com o uso de sistemas de representação de conhecimento é possível categorizar elementos textuais e criar estruturas que auxiliam no processo de indexação e recuperação das bases de documentos. Embora o uso desses modelos possa auxiliar na recuperação de informação clínica, a busca semântica ainda apresenta limitações conhecidas, uma vez que na comunicação humana, significados estão fortemente condicionados à intenção de quem produz e consome informação.

Várias técnicas têm sido propostas e utilizadas para aprimorar a recuperação de informação em PEPs. Contudo, estudos ainda não consideram a relação e influência das intenções na recuperação de informações clínicas. No contexto da área médica, registros em PEPs são predominantemente escritos em linguagem natural e o fato das intenções não serem explicitamente declaradas dificulta o processo de recuperação de informação.

Esse cenário torna desafiador a concepção de mecanismos computacionais que permitam aos profissionais médicos expressarem suas intenções explicitamente, em

[2] https://www.nlm.nih.gov/

particular, levando em consideração esses elementos em algoritmos adequados para selecionar e ordenar a relevância dos documentos em bases de PEPs.

A problemática desta dissertação é considerar intenções como um aspecto diferencial para melhorar os resultados de recuperação de informação em comparação às técnicas de busca sintática e semântica. Para ilustrar o problema em termos concretos, a Tabela 1.1 apresenta dois casos reais de PEPs, um descrito por um clínico geral e outro por um psiquiatra. Os dois prontuários apresentados na Tabela 1.1 contêm a mesma hipótese diagnóstica[3]. Esses casos representam exemplos das dificuldades envolvidas na recuperação considerando as intenções.

Tabela 1.1: Exemplo de PEPs

	Evento	***Clínico Geral***	***Psiquiatra***
1	***Pré Consulta***	Resultado de Exame. Dor nas costas, afebril, dor de cabeça e dor na nunca. PA: 130X90 Peso: 86,500 Nega Alergia Medicamentosa.	Dor no corpo, Dor de cabeça PA: 110X70 Prova do Laço Negativo.
2	***Anamnese***	Mialgia, Artralgia, Febre, Cefaleia e Náuseas. <u>Resultado do Exame</u> Hemograma 07/05/2015 HT 39% Plaquetas 240.000 Leucócitos 3800	Politralgia, Dores Abdominais generalizadas e cefaleia frontal há 24 horas. Nega hemorragias, nega diarreia, nega alterações urinárias, nega vertigem. AP: Relata úlcera gástrica e HAS em uso de captopril 25;1;0;0 Nega alergia medicamentosa. Ao exame: Dentição em péssimo estado de conservação.
3	***Hipótese Diagnóstica***	Dengue [dengue clássico] --- (A90)	Dengue [dengue clássico] --- (A90)
4	***Ação Imediata***	Dispensado.	1) Dipirona 01 ampola; 2) Ranitidina 01 ampola; 3) AD 20 ML, EV Lento.
5	***Prescrição Medicamentosa***	1) Hidratação Oral. 2) 0 l Gatorade + 4,0 l de agua. 3) Repouso. 4) Dipirona.....1fr.	Uso oral: 1) soro caseiro, 3 litros por dia durante 7 dias. 2) Dramim B6 50 mg, 01 comprimido de 8/8 se houver náuseas.

[3] Parte do atendimento médico, voltada à identificação de uma eventual doença.

		5) 40 gts 6/6h se houver dor ou t>37,8. 6) Dexclorfeniramina 2mg.....15cps, 01 cp 6/6h se coceira. 7) Metoclopraminda 10mg.....15 cps 01 cp 8/8h se náuseas.	3) Dipirona 500 mg, 01 comprimido de 6/6 horas se houver febre ou dor. Soro caseiro (para 1 litro): 1 litro de água filtrada ou fervida, 1 colher de chá rasa de sal, 2 colheres de sopa rasas de açúcar.
6	***Exames Laboratoriais***	Nenhum	Hemograma Completo

Considere um cenário onde se deve recuperar a anamnese[4] que tenha o diagnóstico conclusivo de "dengue". O prontuário escrito pelo clínico geral tem a "intenção de prescrever" algo de acordo com o resultado gerado por um pedido de exames que foi solicitado por uma consulta médica anterior. Através dos dados do resultado do exame, é possível determinar se o paciente apresenta ou não sorologia positiva (*i.e.*, está com dengue). Já o prontuário escrito pelo psiquiatra tem a "intenção de descrever" o quadro de saúde do paciente. Nesse momento, ainda existe uma suspeita sobre o diagnóstico e exames laboratoriais são solicitados para comprovar o caso.

Os dois prontuários denotam a intenção do profissional ao reportar o quadro clínico do paciente como segue:

- Prescritiva: O profissional (clínico geral) teve a intenção de prescrever o resultado e os tratamentos diagnosticados em uma consulta anterior. Nesta etapa, as informações sobre o tratamento atingem diretamente a causa e não os sintomas da doença.
- Descritiva: Neste caso, ainda existem dúvidas e, portanto, o médico (psiquiatra) refere o estado emocional do paciente. O tratamento se restringe aos sintomas e não a causa.

Neste cenário de busca (diagnóstico conclusivo) somente a anamnese do clínico geral deveria ser recuperada embora a hipótese diagnóstica de ambos os médicos seja de dengue. O cenário mostra a dificuldade de recuperação de informação em sistemas computacionais da área médica, onde embora existam casos com a mesma hipótese diagnóstica, outras informações médicas como a anamnese diferem significativamente em função da intenção.

[4] Ponto inicial no diagnóstico de uma doença ou patologia.

Este contexto ilustra desafios que mecanismos computacionais devem superar para contribuir com o avanço da recuperação da informação sobre PEPs. Ele apresenta que não apenas as palavras-chave devem ser consideradas de maneira isolada analisando aspectos sintáticos da linguagem, mas também as intenções dos profissionais ao produzir a informação. Nesta investigação, a questão de pesquisa seguinte norteia esta dissertação:

> "Como aprimorar métodos de recuperação de informação textual em prontuários eletrônicos de pacientes pela representação explícita de intenções?"

1.2. Objetivos, Métodos e Contribuições

O objetivo deste trabalho de pesquisa é investigar uma proposta original de recuperação de informação considerando intenções em textos em linguagem natural descritos em PEPs. Segundo Liu (2000), intenções podem ser vistas como atos ilocucionários[5] que representam a unidade básica de comunicação humana. Esses atos transmitem a intenção do falante, *e.g.*, aquilo que a pessoa espera que aconteça, mesmo que implicitamente.

Esta pesquisa assume que métodos e teorias da semiótica e comunicação para a descrição dos aspectos pragmáticos servem como um referencial teórico adequado. Consideramos que elas podem apoiar na formalização e estruturação sobre texto em linguagem natural como meta-dados em prontuários médicos. A hipótese é que o uso de intenções pode melhorar a eficácia dos resultados de busca, aprimorando de fato as medidas de qualidade sobre buscas realizadas pelos profissionais da área da saúde.

A partir do objetivo principal são definidas as seguintes metas de pesquisa:

Meta 1: Investigar um modelo de classificação de intenções a ser utilizado em anotações em documento de linguagem natural;

Meta 2: Definir um algoritmo de busca que considere termos de consultas e a intenção explícita do usuário visando o retorno de documentos médicos relevantes;

[5] Unidade básica na comunicação do ser humano que consiste em conteúdos proposicionais que carregam intenções a ser percebida pelo ouvinte.

Meta 3: Desenvolver um protótipo de sistema de recuperação de informação implementando o algoritmo em um cenário de execução para um estudo de caso;

Meta 4: Realizar avaliações experimentais efetuando comparações do algoritmo proposto em relação aos métodos de recuperação tradicionais.

Como primeiro passo na pesquisa, foram investigados meios para caracterizar as dimensões relacionadas aos tipos de intenções presentes nos prontuários. A partir dessa análise foram extraídos termos recorrentes em PEPs para expressar diferentes classes de intenção. Com base nisto, foi definido um algoritmo original para recuperação de informação que se fundamenta em anotações dos significados dos termos e a intenção em que foram expressos nos documentos médicos. Uma anotação visa definir um meta-dado, que no contexto deste trabalho torna explícito o tipo de intenção e o relaciona a uma sentença de um documento. Investigamos meios de efetuar de maneira assistida as anotações nos prontuários descritos em linguagem natural. O algoritmo foi implementado em um sistema de *software* que foi alimentado com uma base real de aproximadamente 13 mil PEPs.

Após o trabalho de formalização e implementação da solução, validações experimentais avaliaram os benefícios do mecanismo de busca desenvolvido por meio de dois cenários definidos por profissionais da área de saúde. A eficácia da abordagem foi examinada em experimentos que utilizaram como referência um conjunto de documentos médicos de teste previamente analisados pela equipe médica.

Em síntese, esta dissertação efetua as seguintes contribuições:

- Desenvolve um estudo com uma análise que esclarece os tipos de intenções mais relevantes no domínio de PEPs;
- Define um método de recuperação de informação com técnicas de anotação de significados e intenções como meta-dados que informam um algoritmo de ordenação de resultados de busca;
- Desenvolve um sistema de *software* que implementa o método com o algoritmo de busca em um contexto médico;
- Conduz uma avaliação experimental considerando uma base real de prontuários médicos e cenários úteis para os médicos para analisar a eficácia

dos conceitos subjacentes da proposta através de medidas tradicionais da área de recuperação de informação.

Na proposta desenvolvida, o método definido não se restringe aos aspectos léxico-sintáticos do conteúdo, podendo recuperar documentos de maneira mais precisa e contextualizada. Do ponto de vista de aplicação, entendemos que este trabalho pode ser benéfico em auxiliar profissionais da área médica em atividades de pesquisa, de gerenciamento, de saúde pública e ao tratamento dos pacientes.

1.3. Organização da Dissertação

Os capítulos restantes desta dissertação estão estruturados da seguinte maneira:

- O **Capítulo 2** apresenta a fundamentação teórica e metodológica descrevendo conceitos básicos sobre PEPs, recuperação de informação, expansão semântica de consultas, Teoria dos Atos da Fala, e pragmática na comunicação humana.
- O **Capítulo 3** desenvolve um levantamento bibliográfico de trabalhos relacionados que visam propor métodos de recuperação de informação sobre PEPs e investiga pesquisas que objetivam abordar o aspecto de intenção na recuperação de informação.
- O **Capítulo 4** define o método de recuperação de informação proposto objetivando alcançar mecanismos de busca que considerem aspectos ligados com a intenção do usuário. O algoritmo de busca é elaborado nesse capítulo.
- O **Capítulo 5** reporta sobre o sistema de *software* desenvolvido. Apresentamos a arquitetura do *software* descrevendo os componentes e tecnologias utilizadas na construção do protótipo.
- O **Capítulo 6** detalha a avaliação experimental descrevendo os cenários de aplicação no qual a proposta foi testada, as medidas de avaliação e os resultados obtidos.
- O **Capítulo 7** finaliza esta dissertação sintetizando as contribuições científicas obtidas, realça os principais resultados para a área de recuperação de informação e da informática médica. Trabalhos futuros e desafios ainda em aberto são descritos.

Capítulo 2

Referencial Teórico-Metodológico

Neste capítulo definimos os conceitos que fundamentam a solução da proposta na dissertação. A Seção 2.1 apresenta os princípios relacionados com PEPs. A Seção 2.2 descreve fundamentos e avanços gerais sobre mecanismos de recuperação de informação. A Seção 2.3 apresenta o referencial teórico adotado nesta dissertação descrevendo a Teoria dos Atos de Fala e a estrutura de classificação de ilocuções.

2.1. Prontuários Eletrônicos de Pacientes

O médico grego Hipócrates (460 a.C. - 370 a.C.), é considerado uma das figuras mais importantes da história da medicina. Desenvolveu teorias, estudos e constatações de epidemias e suas relações com o tempo. Uma das suas principais teorias se baseava no fato de que as informações médicas deveriam ser armazenadas e agrupadas em ordem cronológica e que deveriam refletir exatamente o curso da doença indicando suas possíveis causas. Sua ideia deu origem ao que chamamos hoje de registros médicos (Totelin, 2006).

O desenvolvimento das ciências médicas fez com que ao longo dos anos, informações fossem acumuladas em grande volume de papéis com uma infinidade de acontecimentos e situações sobre a saúde do paciente. Surgiram diversas dificuldades, dentre elas, anotações ilegíveis, desorganização, perdas, rasuras entre outras, que resultam em prejuízos aos pacientes. Isto tem motivado pesquisadores a investigarem meios de desenvolver mecanismos computacionais que possam minimizar estes tipos de problemas.

Em meados da década de 60 surgiram os primeiros sistemas de informação na área médica com a finalidade de armazenar e organizar dados sobre o histórico de saúde do paciente (Weed, 1968). Os registros médicos passaram então a fazer parte preliminarmente de uma nova organização de armazenamento que consiste em PEPs. Eles compõem um emaranhado de informações sistematizadas e organizadas em formato digital.

Durante os últimos 25 anos, muitos registros médicos foram convertidos para o formato digital com a finalidade de aprimorar a qualidade do tratamento médico e para evitar erros de assistência médica ao paciente. Este cenário tem motivado países como o Brasil a implantar gradativamente sistemas informatizados em unidades públicas de saúde. Já na esfera da saúde privada, *e.g.*, consultórios médicos particulares, clínicas especializadas em saúde, hospitais entre outros, foram impostas sanções pelo governo com o propósito de incentivar a transformação das informações médicas para o formato digital.

O Conselho Federal de Medicina (CFM) define o PEP como: *"Um documento único que constitui um conjunto de informações, sinais e imagens registradas. Os dados são gerados a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo"* (Resolução CFM n. 1.638/2002, art. 1).

No contexto atual, organizações como a fundação *openEHR*[6] e a *Health Level Seven International* (*HL7*)[7] propõem padrões internacionais para a especificação e o uso de PEPs. Para Laforest & Tchounikine (1999), o PEP é a principal ferramenta para a centralização e coordenação de pesquisas médicas. São multidisciplinares, preenchidos e acessados por vários profissionais e não são modificáveis por se tratar de diagnóstico de pacientes. Isso reduz a concorrência dos tratamentos médicos, principalmente em relação às doenças crônicas e psiquiátricas, que por sua vez, podem advir de uma estrutura hereditária que muitas vezes não tem um tratamento completo.

Nesse cenário, o *openEHR*[8] consiste em uma plataforma aberta, orientada ao domínio, para o desenvolvimento de sistemas computacionais no domínio da saúde. A fundação *openEHR* definiu um conjunto de especificações que permitem o armazenamento, gerenciamento e integração de dados em sistemas da área de saúde. O conjunto de especificações e arquitetura do *openEHR* (Heard & Beale, 2007) define padrões de referência para as informações médicas utilizando terminologias externas

[6] http://www.openehr.org/
[7] http://www.hl7.org/
[8] http://www.openehr.org/pt/what_is_openehr

como *SNOMED CT*[9], *LOINC*[10] *e ICD-X*[11]. A arquitetura do *openEHR* viabiliza a construção de um repositório de PEPs em uma abordagem onde dados e modelos de dados são totalmente separados.

Os modelos são conhecidos como "arquétipos" e podem ser compartilhados e reutilizados em diferentes projetos. O *Clinical Knowledge Manager* (CKM)[12] permite recuperar arquétipos, bem como visualizar a documentação padrão associada. Para o conceito de anamnese, por exemplo, existe um arquétipo que a define como um histórico clínico do paciente registrado ou reportado diretamente por ele ao médico. O arquétipo define a anamnese como uma narrativa aberta e/ou associada a uma estrutura com *slots* específicos, *e.g.*, para questões, sintomas e eventos de saúde.

A principal função computacional dos arquétipos é servir de base para consultas de dados nos sistemas. Elas podem ser expressas em uma linguagem baseada em *Structured Query Language* (*SQL*) e *W3C XPaths* extraídos dos arquétipos. Nesse sentido, o *openEHR* contribui na construção de "consultas portáteis" que usam modelos de conteúdo (baseados em arquétipos), e não apenas em esquemas de banco de dados.

Iniciativas como o *openEHR* são certamente chaves para padronizar dados médicos, melhorar a rastreabilidade dos dados (ao gerar parte do *software* a partir de modelos) e consequentemente apoiar recuperação de informação na área médica. Entretanto, elas focam majoritariamente nos aspectos quantitativos, *i.e.*, na recuperação da informação diante de um padrão de arquétipos pré-estabelecido. Os aspectos qualitativos da informação sobre o paciente (do ponto de vista clínico), como o contexto do quadro clínico do paciente e o seu diagnóstico, são constituídos por textos livres em linguagem natural em muitos casos.

No arquétipo relacionado com a anamnese, por exemplo, boa parte da informação qualitativa está em texto livre, com base no relato do paciente. Nele o conteúdo pode ser apenas (e de maneira optativa) parcialmente estruturado. Entretanto, frequentemente, é necessário obter informações precisas que não estejam limitadas quantitativamente como, por exemplo, o número de vezes que o paciente realizou

[9] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[10] https://loinc.org/
[11] http://www.who.int/classifications/icd/en/
[12] http://www.openehr.org/ckm/

aferições de pressão no decorrer do seu histórico. Para obter averiguações clínicas[13] com maior efetividade são necessários dados que informem os valores muitas vezes subjetivos que são obtidos com o paciente; ou seja, informações que auxilie de fato o profissional de saúde. Sob tal ótica, textos livres presentes em PEPs podem fornecer informações relevantes para subsidiar pesquisas na área medica como, por exemplo, a evolução de doenças crônicas.

Assumimos que este trabalho lida com PEPs bem estruturados de acordo com as técnicas atuais, e assim focamos em técnicas de recuperação de informação sobre os textos livres em linguagem natural (*e.g.*, descrições de pré-consulta e anamnese). Mais especificamente, esperamos aprimorar a relevância de buscas de elementos textuais "internos" a esses modelos.

2.2. Recuperação Semântica de Informação

Sistemas atuais armazenam um grande volume de documentos. De maneira geral, os mecanismos de buscas são constituídos por três componentes: (1) base de dados contendo índices para documentos, (2) sentença de busca (consulta) e (3) um algoritmo capaz de determinar a ordenação e como os resultados de busca serão exibidos ao usuário final.

O fato das informações serem usualmente armazenadas sem relacionamentos explícitos dificulta a execução de técnicas que recuperação de informação. Isso faz com que os usuários realizem diversas tentativas de recuperação para encontrar o conteúdo que realmente necessitam (Jenice & Kurian, 2012).

Dias & Santos (2001) explicam que essa dificuldade pode ser entendida pelo fato de que os mecanismos de busca são principalmente projetados com técnicas que exploram comparações léxico-sintáticas em cadeia de caracteres do conteúdo disponível. As principais propostas organizam e agregam documentos textuais em domínios através de ocorrências de palavras/termos contidos nos documentos. Essas técnicas não são usualmente capazes de retornar resultados relevantes devido à complexidade introduzida por termos com polissemia[14] e palavras sinônimas[15].

---

[13] Similar a mapeamento clínico: exame detalhado sobre o histórico do paciente para a identificação ou acompanhamento de eventuais alterações (positiva ou negativa) de saúde do paciente.

[14] Uma palavra assumir mais de um sentido (significado) para além do seu sentido original.

Uma forte necessidade tem sido considerar a estrutura e os significados dos dados na recuperação de informação. Para esse fim, os mecanismos de busca necessitam ter uma representação sobre o conhecimento do domínio quando seus conteúdos são analisados. Isto, por sua vez, exige que a máquina "interprete" o que significa o conteúdo, *i.e*, sua semântica (Dos Reis, Bonacin & Baranauskas, 2014). Técnicas para a representação de conhecimento têm sido investigadas há décadas na literatura em pesquisas na área de Inteligência Artificial.

A semântica é a ciência que estuda os significados e a interpretação do significado de uma palavra, de um signo, de uma frase ou de uma expressão em um determinado contexto. Para Morris (1937), a sintaxe se diferencia da semântica pelo fato de se preocupar com o formato e a formalização dos termos e não com seus significados e interpretações. Os mecanismos de busca investigados (que visam considerar a semântica) têm sido amplamente utilizados para promover desambiguações em termos de polissemias e sinônimos com o objetivo de obter o sentido real sobre os conteúdos consultados pelo usuário (Dos Reis, Bonacin & Baranauskas, 2014).

O processo de recuperação semântica de informação pode ser constituído de três fases principais, como uma abordagem, sendo: (1) os dados são interpretados em linguagem natural, onde há uma extração da relevância dos conceitos da sentença; (2) o conjunto de conceitos são usados para construir consultas que são confrontadas com artefatos que representam o conhecimento do domínio; e (3) finalmente, os resultados são apresentados aos usuários na interface do sistema computacional (Hildebrand, Ossenbruggen & Hardman, 2007).

Para que a representação do conhecimento semântico se concretize, é preciso que os dados sigam padrões e protocolos que garantam interoperabilidade entre sistemas. O desafio de conseguir atribuir um significado (sentido) aos conteúdos publicados de modo que seja perceptível tanto pelos humanos como por computadores (agentes inteligentes) tem sido o foco principal dos estudos da *Web* Semântica (Jenice & Kurian, 2012).

Guha, McCool & Miller (2003) definem recuperação semântica de informação como uma aplicação da *Web* Semântica. Ou seja, são aplicações que fazem uso da

[15] Palavras que possuem significado idêntico ou semelhante a outras palavras.

tecnologia da *Web* Semântica para recuperar informação de maneira mais eficaz ao considerar o significado dos dados disponíveis.

2.3. Expansão de Consulta para Recuperação Semântica de Informação

Com técnicas de expansão de consultas é possível detectar e explorar termos relacionados às palavras-chave originalmente utilizadas pelo usuário. Expansão de consulta pode desenvolver um papel relevante, pois o vocabulário do usuário para um tópico de consulta geralmente é menos diversificado do que o vocabulário do domínio (Chawla & Bedi, 2008). Adicionalmente, os resultados das consultas frequentemente não refletem as necessidades dos usuários devido ao impacto de diversos fenômenos da linguagem que dificultam o processo de busca, tais como ambiguidade de termos, polissemia e homonímia.

Uma maneira de superar esse impacto é pela reformulação das consultas que incluem outros termos relevantes (*e.g.*, termos mais específicos ou genéricos). A Figura 2.1 apresenta um processo de expansão de consulta que tem como finalidade adicionar novos termos relevantes (elemento "*Expansão de Consulta*" na Figura 2.1) na consulta inicial proposta pelo usuário (elemento "*Consulta*" na Figura 2.1).

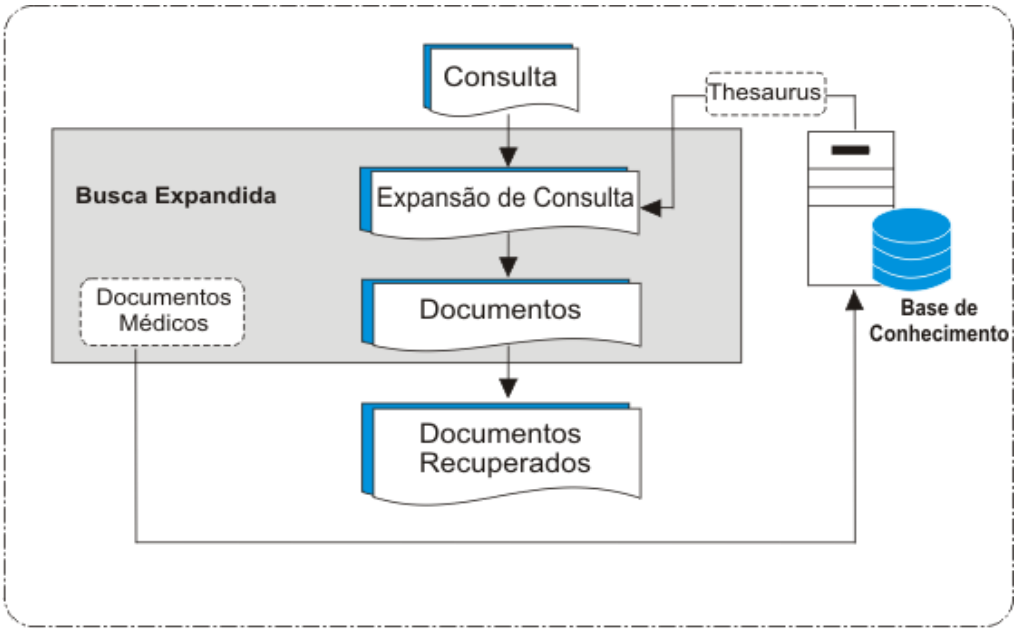

Figura 2.1: Processo de expansão de consulta (Adaptado de Deim, Chavallet & Thuy 2007)

Para Jaladi & Borugerdi (2008), o processo de adição de termos da consulta expandida pode ser manual ou automático. A adição manual baseia-se na experiência do

usuário em inserir novos termos significativos ao contexto. No caso da expansão automática, as ponderações (ocorrência de termos nos documentos pesquisados) devem ser calculadas para todos os termos. De maneira geral, os resultados que tiverem as maiores ponderações são adicionados à consulta inicial.

Sharef & Madzin (2012) argumentam que a expansão de consultas pode estender a consulta original utilizando representações de conhecimento extras, com base em redes semânticas. Isso torna os resultados mais significativos no que diz respeito aos documentos do domínio. Essencialmente, o processo se beneficia de bases de conhecimento, como *thesaurus*, para selecionar novos termos que compõem a consulta. Esses *thesauri* possuem a definição de nomes/rótulos alternativos para refinar os significados, bem como relacionamentos entre diferentes conceitos (Chawla & Bedi, 2008).

De forma complementar, um *metathesaurus* define uma estrutura capaz de agregar conceitos de várias fontes, tais como vocabulários e *thesauri*. Ele também pode associar esses conceitos com descrições semânticas unificadas. Nesta dissertação, exploramos o UMLS como modelo de representação de conhecimento para permitir explorar relações semânticas presentes no domínio médico e, assim, possibilitar o uso de técnicas de expansão de consulta.

2.4. UMLS como artefato de representação de conhecimento médico

O UMLS é um projeto desenvolvido pela NLM (*National Library of Medicine dos Estados Unidos*) desde 1986[16]. Ele consiste de uma extensiva biblioteca terminológica que resulta em uma combinação de aproximadamente 200 bases (fontes) com vocabulários e padrões que representam o conhecimento biomédico em várias línguas. O fato de possuir bases em língua portuguesa possibilita o uso do UMLS no contexto dos PEPs utilizados nesta dissertação.

A Figura 2.2 apresenta um exemplo de consulta[17] no UMLS para o termo "Coração". Nessa consulta são apresentadas as diferentes bases em várias línguas, assim como os conceitos relacionados à "Coração". Por exemplo, no *thesaurus* MeSH

[16] https://www.nlm.nih.gov/research/umls/
[17] Consulta realizada utilizando o *UMLS Terminology Services*: https://uts.nlm.nih.gov/

(*Medical Subject Headings*[18]) o conceito "Coração" está relacionado com outros 24 (parte de cima da Figura 2.2), *e.g.:* "Sistema Cardiovascular", "Endocárdio", "Coração Fetal", "Átrios do Coração", entre outros. Para cada um desses conceitos, são armazenados seus identificadores e características chave como idiomas disponíveis. Cada conceito no UMLS tem um CUI (*Concept Unique Identifier*), que é um código único de identificação que representa o significado no domínio. O CUI é mapeado em códigos de conceitos nas diferentes bases que compõe o UMLS, por exemplo: uma doença no UMLS tem um CUI associado, que por sua vez é mapeado para códigos do CID (Classificação Internacional de Doenças) em suas diferentes versões (*e.g.*, CID9 e CID10).

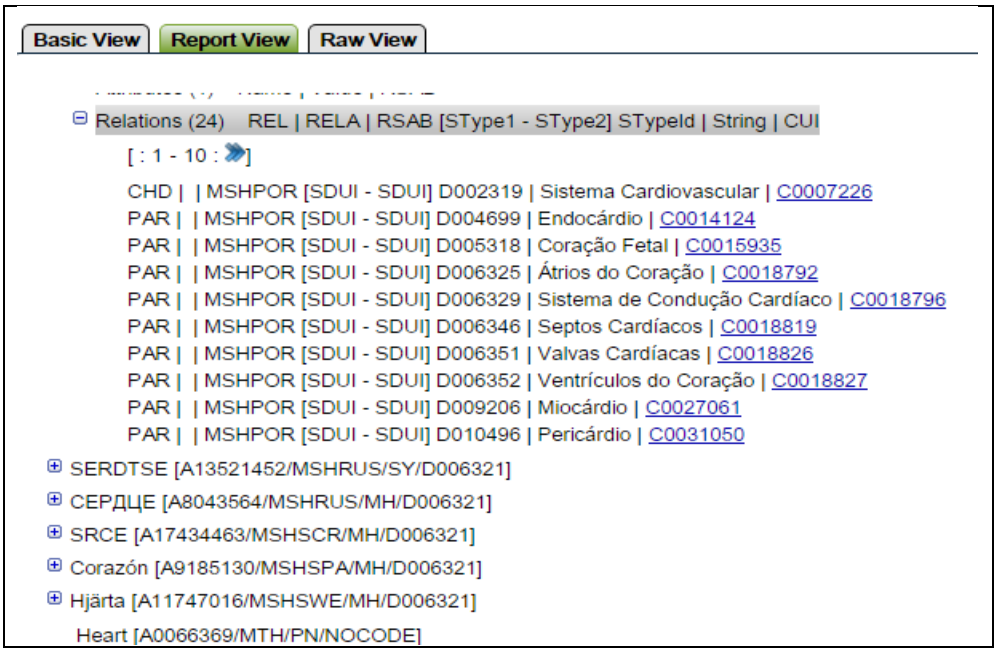

Figura 2.2: Exemplo de Consulta no UMLS

O UMLS possui três componentes que tem a finalidade de formar uma estrutura unificada para diferentes fontes de conhecimento na área biomédica. A Figura 2.3 apresenta um modelo de representação utilizado pelo UMLS onde:

- **Rede semântica** (Camada superior na Figura 2.3): é a classificação de conceitos em tipos semânticos e estabelece relações entre eles. Adicionalmente, fornece uma categorização de tipos semânticos consistente de todos os conceitos representados no *Metathesaurus* do UMLS. As duas

[18] http://www.ncbi.nlm.nih.gov/mesh

principais categorias da rede semântica são entidades (*Entity*) e eventos (*Event*). Por exemplo, todos os antibióticos existentes no UMLS estão categorizados na categoria "*Antibiotic*" que é um tipo específico de "*Pharmacologic Substance*" que é uma "*Chemical Viewed Functionally*", que é "*Chemical*", que é "*Substance*", que por sua vez é uma entidade.

- ***Metathesaurus*** (Camada do meio na Figura 2.3): é o maior componente do UMLS. Ele agrega conceitos de várias fontes (bases), contento assim todas as informações e definições de conceitos do UMLS. Ele é organizado por CUIs que estabelecem relações entre termos semelhantes em aproximadamente 200 bases distintas.

- **Bases de conhecimento** (Camada do inferior na Figura 2.3): conjunto de vocabulários controlados e padrões internacionais com informações lexicográficas. Elas incluem resultados de esforços de categorização e padronização de termos médicos que são explorados a décadas como o CID que teve origem no século 19. Outros exemplos de bases amplamente conhecidas na área médica são: MeSH, SNOMED CT, NCI e LOINC.

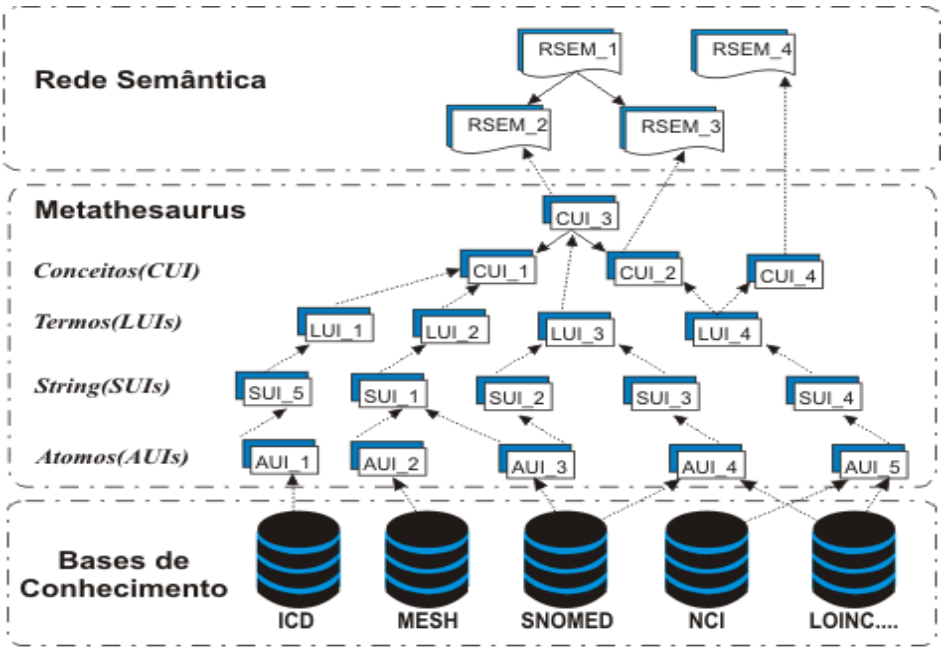

Figura 2.3: Estrutura da UMLS – (Adaptado de Deim, Chavallet & Thuy 2007)

A representação de conceitos nas três camadas leva a um nível mais alto de abstração do conteúdo de um texto médico. Esta abordagem orientada a conceitos é

interessante para métodos computacionais que tem a finalidade de aumentar a correspondência entre termos do domínio considerando sinônimos, abreviações e variações dos termos (Deim, Chavallet & Thuy, 2007).

Ao utilizar o UMLS como parte de uma solução para a expansão semântica de consultas de busca (c*f.*, seção 2.3) podemos contar com a terminologia modelada nas principais bases existentes em uma visão única. Uma palavra-chave expressa em uma consulta do usuário pode ser consultado no UMLS para recuperar sua terminologia padrão, os conceitos relacionados e tipos semânticos associados.

Por exemplo, um usuário que consulte "Coração" em um mecanismo de busca com PEPs pode estar também interessado em PEPs que contenham termos relacionados em suas descrições, como por exemplo, um PEP que relata problemas no "endocárdio" e "átrios do coração".

2.5. Semiótica, Teoria dos Atos da Fala e Classificação de Ilocuções

Semiótica é a ciência que estuda os processos de significação e o conceito de signo. Trata-se de uma área que está intrinsecamente ligada à comunicação ou linguagem como um meio de representação dos fatos (Santaella, 2004). Ela investiga todos os meios pelo qual o homem se comunica, considerando a forma verbal ou não verbal.

Bonacin (2004) destaca que a semiótica tem sido objeto de estudo de pesquisadores em Linguística, Estudos de Mídia, Ciências Educacionais, Antropologia, Filosofia da Linguagem entre outros. Peirce (1931-1958) afirma que existem três campos distintos de estudo da semiótica: a Sintaxe, a Semântica e a Pragmática.

De acordo com Morris (1937), a *Sintaxe* lida com a relação entre os signos. A *Semântica* lida com as relações entre signos e os objetos a que eles se referenciam (o que o signo denota e designa); enquanto a *Pragmática* lida com a relação entre os signos e as pessoas que os interpretam (o que o signo expressa).

A *Pragmática* define o comportamento intencional, onde o contexto é utilizado para indicar intenções. Portanto, o elemento central deste trabalho é considerar aspectos pragmáticos como meio para aprimorar mecanismos de recuperação de informação, indo além de técnicas que consideram a sintática e a semântica.

Além do conceito de pragmática definido no contexto da Semiótica, este trabalho se fundamenta na Teoria dos Atos de Fala. Essa teoria surgiu em meados dos anos cinquenta, tendo como pioneiro o inglês Austin (1976), seguido por Searle (1969).

A linguagem era entendida apenas para descrever o estado das coisas distinguindo-os em proferimentos *constatativos* que descreve o estado das coisas, podendo ser verdadeiro ou falso e *performativos* como enunciado que não afirma e nem nega, mas realiza um ato quando é pronunciado. Austin colocou em questão essa visão, mostrando que certas afirmações não servem para descrever nada, mas sim para realizar ações.

Conforme apresenta Ilari (2003), os atos *constatativos* são contrários aos atos *performativos*, pois relatam as coisas do mundo que podem ser analisados sob o ponto de vista da verdade ou falsidade. Por exemplo, na frase "*O paciente possui sorologia positiva de Dengue",* trata de uma afirmação, relato ou descrição que pode ser avaliada sob critérios de veracidade ou falsidade.

Partindo desse pressuposto, uma nova visão sob o prisma dos atos *constatativos* foi formulada. As frases que relatavam algo, ou seja, que somente possuíam a função de informar passaram a ser analisadas como frases em que as ações verbais indicavam ações praticadas no momento da fala. Sentenças, como por exemplo: "*Diante do quadro de sorologia positiva da Dengue, recomendo repouso de 10 dias a contar do dia de hoje*", descrevem não apenas o fato de serem verdadeiras ou falsas, pois ao serem proferidas em circunstâncias apropriadas, levam à execução de uma ação ou prática, não se atendo apenas à descrição ou declaração.

Um ato de fala, ou seja, uma elocução que tem função performativa, é considerado por Austin a unidade básica de significação. Segundo Austin, atos de fala são constituídos por três dimensões, respectivamente:

1. Ato locutório: Corresponde ao ato de pronunciar um enunciado, *i.e.*, nas palavras e sentenças empregadas de acordo com regras gramaticais.

2. **Ato ilocutório**: Corresponde ao ato que o locutor realiza quando pronuncia um enunciado em certas condições comunicativas e com certas intenções, tais como ordenar, avisar, criticar, perguntar, convidar, ameaçar, *etc*. Assim, num ato

ilocutório, a intenção comunicativa de execução vem associada ao significado de determinado enunciado.

3. Ato perlocutório: Diz respeito aos efeitos que um dado ato ilocutório produz no alocutário[19]. Efeitos como convencer, persuadir ou assustar descrevem este tipo de ato de fala, pois correspondem ao efeito causado no alocutário.

Para Searle (1969 & 1976) os atos ilocutórios são classificados da seguinte maneira:

1. Ato ilocutório assertivo: Corresponde a uma afirmação de como o mundo é, carrega comprometimento com o valor relativo de (verdade ou falsidade), exemplos ligados a este tipo de atos são: reivindicações, informações previsíveis entre outros.

2. Ato ilocutório diretivo: atos de fala a partir dos quais o falante pretende levar o ouvinte a fazer coisas. Estão-lhe associados verbos como convidar, pedir, requerer, ordenar.

3. Ato ilocutório compromissivo: Este ato gera o compromisso ou a intenção de realizar algo ou uma ação no futuro. Fazem parte deste ato palavras como promessas, com a intenção de se comprometer a realizar uma determinada ação no futuro.

4. Ato ilocutório expressivo: Este ato é empregado em frases em que o locutor pretende expressar, atitudes ou emoções como, por exemplo, pedir desculpas, agradecimentos, felicitações etc.

5. Ato ilocutório declarativo: É usado quando o falante tem a intenção de mudar o mundo, instituindo ou alterando o estado das coisas através de expressões vocais.

Com base na Teoria Semiótica e da Teoria dos Atos de Fala, Liu (2000) desenvolveu um arcabouço teórico para representar elementos de intenção. Segundo Liu (2000), a comunicação é um sistema intencional em que os seres humanos agem e

[19] Alocutário: pessoa a quem o locutor dirige um ato de fala

interagem uns com os outros para conseguir os objetivos que podem estar relacionados com a comunidade ou seus indivíduos.

Nesse sentido, para que haja comunicação, é preciso conter dois agentes, o locutor e o receptor que são responsáveis pela emissão, recepção e consequências da comunicação que são expressos através dos atos de fala. Contudo, o simples fato de proferir um enunciado em meio a uma comunicação não garante a sua realização.

Para que um enunciado seja bem-sucedido, para que a ação por ele designada seja de fato realizada, é preciso, ainda, que as circunstâncias sejam adequadas. Um enunciado pronunciado em circunstâncias inadequadas não é falso, mas sim nulo e sem efeito.

Uma vez que os atos ilocutórios são os atos que caracterizam as intenções nas expressões dos locutores, eles possuem a capacidade não só de representar relevância do mundo social, incluindo as suas ações, mas também fazem parte de suas ações (sociais). Por conseguinte, informação comunicada através de linguagem tem seu complemento no mundo social de situações, instituições e padrões culturais.

Para Liu (2000) um ato de comunicação, definido com base em Atos de Fala, pode ser definido como uma estrutura ternária consistindo em *locutor*, *receptor* e a *mensagem*. A mensagem pode ainda ser dividida em duas partes, o *conteúdo* e a *função* (Liu, 2000). Na Figura 2.4, a linha pontilhada representa o ato de comunicação entre o locutor e o receptor; nota–se que a comunicação possui duas vias, a primeira do locutor para o receptor e a segunda do receptor para o locutor. Já a linha sólida representa a intenção e a reação do enunciado proferido.

O *conteúdo* agrega o significado da mensagem expresso na proposição. O significado e a interpretação da mensagem dependem do ambiente no qual ela é "elocucionada", do âmbito social ou do comportamental humano, tanto do locutor quanto do receptor. A *função* especifica as ilocuções (atos da fala) através de signos, refletindo a intenção do locutor na mensagem. Inspirado na Teoria dos Atos de Fala e fundamentado na Teoria da Semiótica, Liu (2000) propôs um arcabouço lidando com diversas dimensões para classificar as ilocuções.

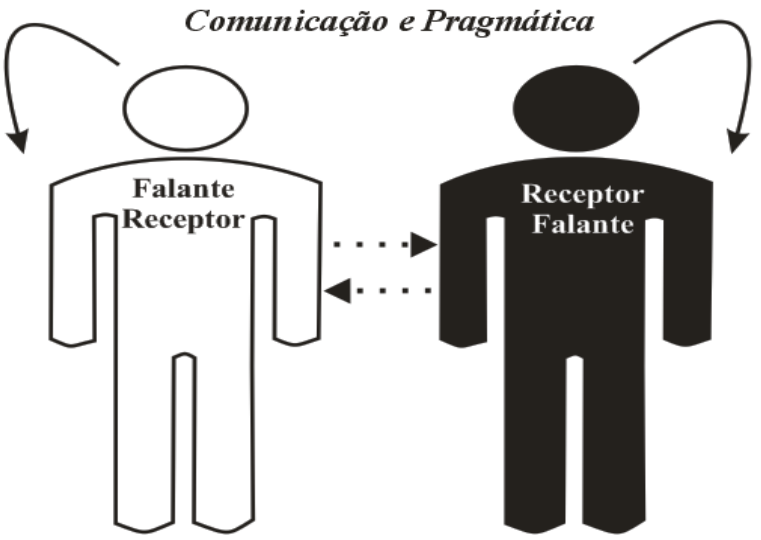

Figura 2.4: Modelo de representação de comunicação, adaptado de Liu (2000).

As ilocuções podem ser agrupadas em três dimensões. Em uma das dimensões faz-se a distinção entre "invenções" descritivas e prescritivas, outra "modos" afetivos e denotativos e finalmente em diferentes "tempos", futuro ou presente/passado. Se uma ilocução, em um ato de comunicação, tende a expressar o estado emocional do locutor esta é classificada como afetiva, senão ela é denotativa.

Se uma ilocução tem função de expressar um efeito inventivo ou instrutivo, esta é uma ilocução prescritiva, senão ela é descritiva. A classificação do tempo é feita a partir do efeito social que uma ilocução causa no ato de comunicação.

Os agrupamentos dessas dimensões dão origem à estrutura de classificação de ilocuções (Liu, 2000), onde as *funções* nos atos da fala são representadas por meio de um "cubo" (Figura 2.5). No total, oito classes foram definidas. Por exemplo, a ilocução do tipo "Proposta" está no futuro segundo a dimensão tempo, denota uma prescrição e está no modo denotativo.

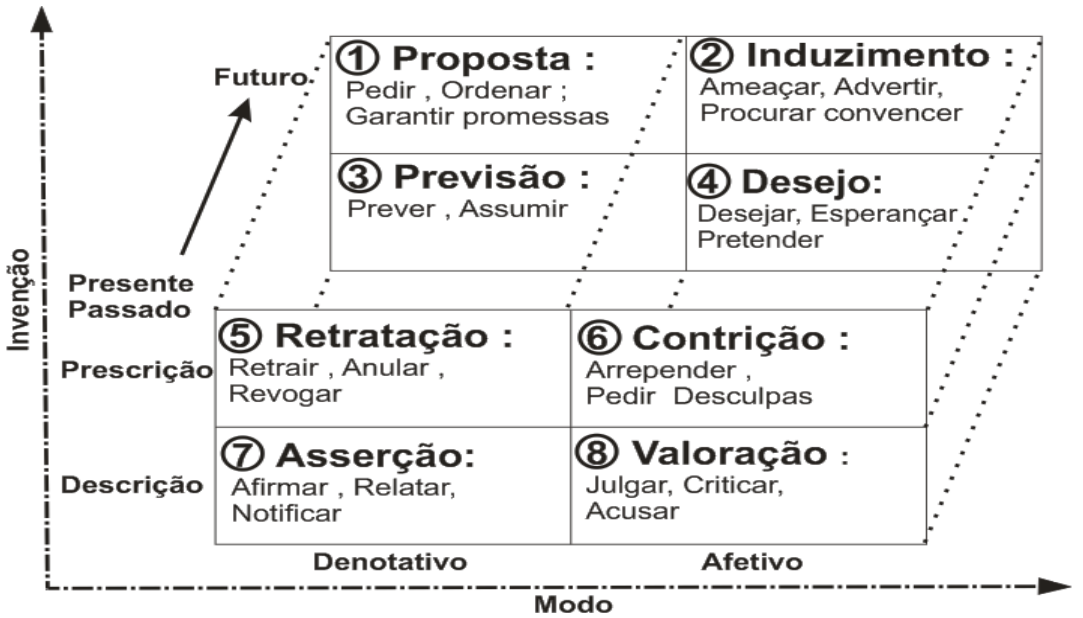

Figura 2.5: Estrutura de Classificação de Ilocuções, Adaptado de Liu (2000).

A Tabela 2.1 apresenta exemplos de mensagens extraídas de conteúdo de PEPs (casos reais escritos por médicos) em uma UPA que evidenciam explicita ou implicitamente os diferentes tipos de ilocuções da Figura 2.5.

Tabela 2.1: Exemplos de mensagens no contexto médico por tipos de ilocução.

Tipo de Ilocução	Mensagem Exemplo
1- Proposta	"Tome este remédio de oito em 8 horas durante sete dias."
2- Indução	"Encaminho o paciente para o urologista por ser um caso específico da área e por necessitar de procedimentos especializados. O paciente deve procurar o mais breve possível devido à suspeita de agravamento do quadro clínico."
3- Previsão	"Paciente deu entrada no pronto atendimento, Hipótese Diagnóstica: Outras conjuntivites virais --- CID-- (B308) ", previsão de retorno de 7 dias a contar da data de hoje para avaliação clínica."
4- Desejo	"Paciente deu entrada na consulta sem apresentar quadro clínico de enfermidade. O mesmo relata que a obesidade tem atrapalhado suas atividades e gostaria de um encaminhamento para cirurgia de estômago."
5- Retratação	"Retorno da consulta após 15 dias com diagnóstico Dengue [dengue clássico] --- (A90), onde o paciente relata ter seguido corretamente o tratamento proposto. Seus exames não apresentam divergências nas plaquetas, portanto suspendo o tratamento."

6- Contrição	"Em relato, paciente diz sofrer de depressão, diz também que fazia tratamento com médico psiquiatra, porém parou há três meses. Diz que se arrepende de ter parado e que está se sentindo mal, suas angustias aumentaram. Hipótese diagnostica: Episódio depressivo leve --- (F320)."
7- Asserção	"Paciente deu entrada no pronto atendimento após ter sofrido um acidente de moto. Os primeiros cuidados médicos foram providenciados e o relatório dos procedimentos realizados estão junto com o prontuário. Encaminho para o hospital para RX da clavícula. Hipótese Diagnóstica: Fratura da clavícula --- (S420)."
8- Valoração	"Paciente em tratamento e avaliação de asma há sete dias. Evolui em melhoria satisfatória do quadro. Ao exame: REG: Taquidispinéico, alerta, reativo, febril, acianótico. MV diminuindo globalmente com estertores crepitantes em base direta."

2.6. Síntese do Capítulo

Neste capítulo foi apresentada a motivação e relevância dos prontuários médicos eletrônicos na área de informática médica, e descrito os conceitos e avanços fundamentais de métodos de recuperação semântica de informação. Nossa proposta reutiliza técnicas em busca semântica em um mecanismo de recuperação de informação que leva em consideração as ilocuções. Para tanto, definimos a técnica de expansão de consultas semânticas e apresentamos o UMLS que será explorado com parte da solução.

Nesta pesquisa, visamos considerar explicitamente os tipos de ilocução como meio de representar e estruturar intenções em textos descritos em linguagem natural do domínio médico. O objetivo é levar em conta a influência dos tipos de ilocução em um algoritmo que recupere de maneira mais criteriosa o conjunto de resultado de busca.

Para esse fim, descrevemos as teorias que são o quadro de referência para o desenvolvimento desta dissertação, que inclui a estrutura de classificação de ilocuções que é diretamente explorada no método de recuperação de informação. O próximo capítulo descreve uma revisão da literatura analisando os trabalhos correlacionados a esta pesquisa.

Capítulo 3

Revisão do Estado da Arte

Neste capítulo são apresentados os trabalhos relacionados a esta dissertação por meio de uma revisão da literatura. As pesquisas abrangem trabalhos sobre técnicas e mecanismos de recuperação de informação que consideram em particular intenções. A seção 3.1 detalha o procedimento conduzido para a realização da pesquisa bibliográfica. A seção 3.2 apresenta a síntese e análise dos trabalhos identificados sobre recuperação de informação que propõem meios para considerar intencionalidade. Já a seção 3.3 apresenta uma análise dos trabalhos relacionados à recuperação de informação em PEPs. A seção 3.4 realiza uma discussão e posicionamento sobre o estado da arte.

3.1 Método de Revisão, Seleção e Análise da Literatura

Diversos modelos e técnicas de recuperação de informação têm sido estudados, formando uma vasta literatura sobre o tema. Diante desse cenário, foi realizada uma investigação utilizando palavras-chave escolhidas de acordo com a proximidade com o objetivo desta dissertação. O método adotado que resulta em critérios para a classificação dos artigos relacionados mais relevantes seguiram os seguintes passos:

- **Passo 1:** Pesquisas nas bases científicas utilizando as principais combinações de palavras-chave. As bases científicas consideradas foram: ACM Digital Library, IEEE Xplore, Springer, ScienceDirect e Google Scholar;
- **Passo 2:** Pré-seleção dos artigos com base na análise dos títulos e resumos de acordo com a proximidade de assuntos e problemática abordados nesta pesquisa. Para tanto, foram lidos os títulos e resumos dos primeiros 200 artigos pesquisados por palavras-chave em cada base;
- **Passo 3:** Análise e categorização integral dos artigos pré-selecionados;
- **Passo 4:** Síntese dos resultados, avaliação crítica, exclusão de trabalhos com resultados poucos relevantes e tabulação dos objetivos, pontos fortes e limitações de cada trabalho;

- **Passo 5:** Escolha de trabalhos a serem reportados nesta dissertação considerando a proximidade com o objetivo proposto como principal critério de inclusão e exclusão dos trabalhos na pesquisa. O grau de proximidade foi atribuído após a leitura completa pelos pesquisadores e utilizado para determinar se o artigo deveria ser considerado.

3.2 Recuperação de Informação com base em Intenções

A Tabela 3.1 apresenta uma síntese dos resultados da pesquisa bibliográfica dos artigos localizados nas principais bases de publicações científicas com relação à recuperação de informação com base em intenções. Para tanto foram utilizadas as seguintes combinações de palavras-chave[20]:

A. "*Information retrieval" Intention*

B. "*Information recovery" Intention*

C. "*Information retrieval" Intended*

D. "*Information retrieval" Purpose*

E. "*Information recovery" Purpose*

F. "*Semantic search" Intended*

G. "*Semantic search" Intention*

H. "*Semantic search" Purpose*

A pesquisa bibliográfica foi realizada no mês de agosto de 2015, sem filtro por data de publicação nas bases. As colunas da Tabela 3.1 indicam a quantidade dos resultados encontrados nos passos do método de pesquisa bibliográfica adotado (*cf.* Seção 3.1). As linhas indicam as bases de publicações. Logo após o processo de escolha e análise dos trabalhos, foram selecionados os principais artigos relacionados.

Tabela 3.1: Resultado da investigação bibliográfica com artigos selecionados sobre recuperação de informação baseada em Intenções.

Base de Pesquisa	**Passo 1**								**Passo 2**	**Passo 3**	**Passo 4**	**Passo 5**
	A.	**B.**	**C.**	**D.**	**E.**	**F.**	**G.**	**H.**				
ACM *Digital Library*	6142	40	10860	33107	148	735	421	1751	18	6	6	3
IEEE *Xplore*	194	0	339	1369	11	9	8	21	22	5	5	4

[20] As palavras foram digitadas diretamente nos mecanismos de buscas para evitar o uso de aspas como caracteres especiais

Springer	5347	36	12647	28004	254	959	398	1814	3	3	0	1
ScienceDirect	2775	22	7508	15471	215	297	141	597	12	4	0	0
Google Scholar	54200	901	124000	649000	4000	14700	7050	33500	1	1	1	1

A Tabela 3.2 apresenta os trabalhos resultantes do processo de seleção de artigos considerados os mais significativos conforme resultados da Tabela 3.1 (selecionados no Passo 5). Apresentamos uma síntese dos principais pontos fortes, limitações e referências.

Tabela 3.2: Síntese dos trabalhos relacionados na pesquisa bibliográfica sobre recuperação de informação baseada em intenções

1- Título	***Document Title Patterns in Information Retrieval***
Objetivo	Propõe um mecanismo de recuperação de informação que visa relacionar o título e o conteúdo do documento por meio da representação de intenções.
Ponto(s) Forte(s)	Apresenta um experimento realizado com 4663 documentos com eficácia de aproximadamente 90% em que as palavras-chave dos títulos determinam a intenção contida nos documentos, sendo que o método proposto é independente de um domínio especifico.
Limitações	A proposta é limitada ao considerar apenas a relação entre o título e o resumo dos documentos. A abordagem é baseada em casamento de padrões e identificação de verbos e construções pré-definidas, o que restringe os meios de representar computacionalmente as intenções.
Referência	Montes-y-Gomez & Lopez-Lopez (1999)
2- Título	***Information Retrieval Techniques to Grasp User Intention in Pervasive Computing Environment***
Objetivo	Define um modelo de dados no contexto de recuperação de informação no qual intenções do usuário são detectadas através de sensores em um ambiente de computação pervasiva.
Ponto(s) Forte(s)	Através de sensores, um módulo computacional interpretador obtém a identificação da intenção do usuário no momento da consulta. Explora a classificação de 17 categorias de intenções implementadas no mecanismo de recuperação que permitem uma melhor ordenação dos resultados.
Limitações	Lida superficialmente com termos polissêmicos (*i.e.*, aqueles que possuem vários significados). Isso pode gerar situações que não atendam às necessidades dos usuários no momento da consulta. Por exemplo, uma intenção relacionada ao termo "entrar" pode denotar (1) o sentido de ir a algum lugar ou (2) abrir uma porta. A segunda sentença está relacionada a contextos mais próximos de porta e janela. Logo o mecanismo precisa ser ajustado de acordo com o *status* do objeto de busca. Por fim, necessita-se de um objeto ligado a um domínio para que o interpretador obtenha satisfatoriamente a busca.
Referência	Hwang & Choi (2011)
3- Título	***Search Bot: Search Intention Based Filtering Using Decision Tree Based***

	Technique
Objetivo	Apresenta um modelo de busca por meio de agentes inteligentes e filtros que visa capturar a intenção do usuário, gravar o perfil e retornar resultados conforme a busca por palavras-chave.
Ponto(s) Forte(s)	A técnica das caixas de seleção dos resultados previamente filtrados pelo G*oogle* permite ao algoritmo gravar a intenção (com base nos títulos das buscas). Este método permite detectar qual conteúdo foi relevante diante da solicitação requisitada pela palavra-chave. Neste sentido o usuário pode refazer a consulta com base nos resultados marcados nas caixas de seleção alterando a ordenação dos conteúdos na *Web*. Um experimento foi realizado com 20 consultas dos quais os 20 primeiros resultados foram considerados para análise. O método implementado representou 90,25% de exatidão nas consultas contra 60,25% quando comparado com os resultados tradicionais resultantes do *Google*.
Limitações	A recuperação de informação apenas acontece, e somente se, o usuário selecionar e marcar as caixas de seleção diante dos resultados previamente classificados pelo *Google*. Logo, várias consultas podem ser necessárias para classificar conteúdos de interesse do usuário.
Referência	Gupta & Gupta (2012)
4- Título	***A Web search analysis considering the intention behind queries***
Objetivo	Propõe um algoritmo para categorizar o comportamento das consultas realizadas pelos usuários. O algoritmo visa compreender a intenção do usuário para maximizar a eficácia da busca através de uma análise sobre as palavras-chaves digitadas pelos usuários que são armazenadas.
Ponto(s) Forte(s)	O trabalho apresenta boa correlação entre palavras-chave com as variáveis de comportamento: 1 - "informativas" (pesquisas em que o usuário formula com a intenção de achar um conteúdo específico). 2 - "navegação" (quando o usuário procura encontrar um local específico) e 3 - "transacional" (quando o usuário realiza transações, downloads, compras, *etc*.). Segundo os experimentos realizados que levam em consideração o número de cliques nas consultas, apenas 10% registram mais de 10 cliques para chegar as suas respostas.
Limitações	Os tipos de intenções consideradas pelo método se limitam as variáveis "informativas", de "navegação" e "transacional". A técnica desenvolvida não permite ao usuário declarar explicitamente outras intenções relacionadas ao propósito geral.
Referência	Mendoza & Baeza-Yates (2008)
5-Título	***Ranking of Web Documents using Semantic Similarity***
Objetivo	Desenvolve uma abordagem que considera todas as relações relevantes entre as palavras-chave que exploram a intenção do usuário. A técnica calcula a fração dessas relações em cada página *Web* em uma base para determinar a ordenação da consulta fornecida pelo o usuário.
Ponto(s) Forte(s)	O método proposto apresentou resultados superiores em uma análise de desempenho de correlação dos 50 registros do experimento (combinação de métodos de busca sintática e uso de ontologias) comparado com resultados providos pelo motor de busca *Google*. Foi examinado a similaridade entre a intenção do usuário no momento das consultas e o conteúdo retornado.
Limitações	A técnica de recuperação elaborada limita explorar elementos de intenções apenas aos títulos do documento enquanto que o corpo do conteúdo do documento não é

	considerado. Adicionalmente, o usuário não possui meios de explicitar suas intenções no momento da realização das consultas.
Referência	Chahal & Kumar (2013)
6- Título	***Using social data as context for making recommendations: an ontology based approach***
Objetivo	Define um modelo sobre interesse do usuário que serve como uma interpretação da intenção do usuário que auxilia durante os processos de recomendação ou de recuperação de informação.
Ponto(s) Forte(s)	O trabalho aborda a construção de modelos baseados na intenção do usuário através da criação de classes ontológicas. O uso desse recurso nos mecanismos de busca podem recuperar informações mais relevantes quando aplicado em contexto de redes sociais.
Limitações	A criação das classes ontológicas é realizada de forma automática e é fundamentada no interesse do usuário. Nesse sentido, páginas pesquisadas com conteúdos irrelevantes podem criar perfis não relevantes de usuários.
Referência	Noor & Martinez (2009**)**
7- Título	***Modelling Knowledge with ZDoc for the Purposes of Information Retrieval***
Objetivo	Apresenta um mecanismo de busca para recuperação de informação com base em uma abordagem sintática e semântica em linguagem de textos naturais. A abordagem utiliza grafos conceituais para a representação de conceitos relacionados a um domínio.
Ponto(s) Forte(s)	A concatenação das expressões léxico-sintáticas, conhecimento semântico e representação de aspetos pragmáticos permitem aos usuários explicitarem suas intenções no momento da consulta. Esse processo permite não apenas a derivação de uma estrutura semântica através da base de informação sintática, mas o cálculo direto das relações entre conceitos dentro do grafo que verifica todas as representações entre o agrupamento de palavras e seu conteúdo formalmente especificado.
Limitações	Não foram realizados experimentos significativos que informam a eficácia do mecanismo de busca. Adicionalmente, a representação de conhecimento (grafos conceituais) é geralmente um obstáculo para a formalização em grande escala.
Referência	Zinglé (2006)
8- Título	***Intelligent Ink Annotation Framework that uses User's Intention in Electronic Document Annotation***
Objetivo	Desenvolve um arcabouço através de um quadro de anotação de caneta "Tinta Inteligente". Este artefato visa promover a maximização de aprendizagem dos sistemas de anotação, detectando intenções de comportamento naturais através de anotação em documentos baseados em papel. O arcabouço reconhece "Segmentação de conteúdo" **(variedade de conteúdos disponíveis)** e "Comentários" **(comentários sobre os conteúdos)**, que podem, por exemplo, estarem escritos fora da margem do documento ou até mesmo sublinhados.
Ponto(s) Forte(s)	Os experimentos conduzidos apontam efeitos positivos no uso do arcabouço de anotação inteligente. Este fato se deve a capacidade do protótipo interagir de forma dinâmica com o usuário que pode utilizar dois tipos de anotações: (1) sentenças destacadas pela caneta promovem a abertura automática de uma janela quando encontram conteúdos associados "**Segmentação de conteúdo**" permitindo inserir comentários e intenções sobre as consultas e (2) o sistema permite anotações

	manuais sem exibir conteúdo associados, criação de novos "**Comentários**" que não estejam associados à algum termo contido nos documentos.
Limitações	Experimentos foram realizados com apenas 4 participantes. O método proposto precisa que a caneta complete a sentença do começo ao fim para exibir a "Segmentação de conteúdo" e "Comentários". Sentenças muito longas exigem alta precisão ao utilizar a "caneta inteligente".
Referência	Asai & Yamana (2014)
9- Título	***IntentSearch: Capturing User Intention for One-Click Internet Image Search***
Objetivo	Propõe uma nova abordagem de pesquisa de imagens na internet que requer apenas um clique entrada. Através disso, capturar a intenção do usuário diante de um esquema adaptativo que calcula a semelhança visual com a imagem de consulta.
Ponto(s) Forte(s)	A técnica define que as palavras-chave no momento da busca sejam expandidas conforme a intenção do usuário ao realizar a busca, através da comparação entre semelhança visual e a imagem de consulta. Partindo do princípio de que o usuário não possua conhecimento suficiente sobre a descrição textual da imagem que procura, o método apresenta um conjunto de imagens que possuam semelhança visual ao usuário e ao clicar na imagem, a sentença de busca é expandida trazendo imagens mais relevantes e próximas da intenção do usuário. A avaliação dos usuários teve os seguintes resultados: 40,88% Muito melhor, 26,90% Pouco melhor, 22,58% Similar, 7,82 Pouco pior e 1,82% muito pior.
Limitações	A técnica de expansão das palavras-chave através da seleção de imagem se limita ao contexto principal da imagem. O resultado é restrito caso o usuário tenha a intenção de pesquisar parte da imagem ou, por exemplo, uma imagem de fundo ao conteúdo principal.
Referência	Tang Xiaoou *et al.* (2012)

A pesquisa bibliográfica apontou um número considerável de estudos focados no conceito de "intenções" que reforçam os esforços da área para possíveis melhorias voltadas a recuperação de informação com base em intenções. Os trabalhos categorizados e selecionados (Tabela 3.2) com maior proximidade deste trabalho de dissertação apresentam diversificados campos de investigação. Eles vão desde pesquisas visando a captura da intenção do usuário através das palavras-chave inseridas no título da busca (Montes-y-Gomez & Lopez-Lopez, 1999), até o uso de canetas com sensores que detectam as intenções dos usuários através de anotações em papel (Asai & Yamana, 2014).

Embora essas contribuições sejam relevantes, o conhecimento científico ainda requer avanços na elaboração de métodos computacionais que interpretem o conhecimento considerando o uso de intenções em ambos, na consulta e no conteúdo dos documentos, e que explore de fato esse aspecto na recuperação de informação. A literatura apenas explora intenções em elementos da consulta, por exemplo, ao dar apoio

ao usuário expressar suas intenções na expressão de entrada da busca ou ao tentar inferir qual a intenção do usuário no momento da busca. Em um contexto mais específico, técnicas de recuperação aplicadas aos conteúdos da área de saúde devem considerar os conceitos, a relação entre esses conceitos e a intenção no momento que foram inseridos para melhorar a qualidade da recuperação em PEPs. Assim, neste trabalho, exploramos também a intenção do usuário ao escrever o PEP e comparamos com a intenção declarada na recuperação da informação, o que não foi encontrado na revisão realizada.

3.3. Recuperação de informação em prontuários médicos eletrônicos

Técnicas de recuperação de informação em sistemas de informação médicos se tornam cada vez mais relevantes. Isso inclui dados específicos dos tratamentos de pacientes e informações baseadas no conhecimento científico, que dá ênfase à medicina baseada em evidências (Hambury, 2012).

A Tabela 3.3 apresenta uma síntese do total de artigos encontrados nas principais bases de publicação científicas no que diz respeito à recuperação de informação em prontuários médicos eletrônicos. Para tanto, foram utilizadas as seguintes combinações de palavras-chave[21]:

A. *"Information retrieval" "Electronic medical records"*

B. "*Information retrieval" EHR*

C. "*Information recovery" "Electronic medical records"*

D. “*Information recovery" EHR*

E. "*Information retrieval" "Medical information"*

F. "*Information recovery" "Medical information"*

G. "*Semantic search" EHR*

H. "*Semantic search" "Information medical"*

A pesquisa bibliográfica foi realizada no mês de agosto de 2015, sem filtro por data de publicação. As colunas da Tabela 3.3 indicam a quantidade de resultados encontrados nos passos do método empregado (conforme descrito na subseção 3.1). As linhas indicam as bases de publicações pesquisadas.

[21] As palavras foram digitadas diretamente nos mecanismos de buscas para evitar o uso de aspas como caracteres especiais

Tabela 3.3: Resultado da investigação bibliográfica com artigos selecionados sobre Recuperação de Informação na área médica.

Base de Pesquisa	Passo 1								Passo 2	Passo 3	Passo 4	Passo 5
	A.	B.	C.	D.	E.	F.	G.	H.				
ACM *Digital Library*	245	245	1	2	1457	3	28	2	28	7	3	1
IEEE *Xplore*	22	30	0	0	588	3	1	0	31	8	4	3
Springer	291	223	1	1	1038	2	16	0	23	3	1	0
ScienceDirect	269	192	2	3	781	5	15	1	26	5	5	3
Google Scholar	3080	5110	36	84	16200	80	478	48	20	1	1	1

A Tabela 3.4 apresenta os artigos mais relevantes selecionados em nossa pesquisa e relacionados à recuperação de informação em bases de PEPs. Descrevemos, brevemente, os objetivos de cada trabalho selecionado no Passo 5, seus pontos fortes e limitações.

Tabela 3.4 –Síntese dos trabalhos relacionados com recuperação de informação em prontuário médico eletrônico

1- Título	***Intuitive justifications of medical semantic search results***
Objetivo	Propõe a construção de um mecanismo de busca semântica que avalia a inteligibilidade das consultas. Por exemplo, em muitos casos um profissional médico pode ter dificuldades para explicar a uma criança que ela foi diagnosticada com uma doença grave. No sentido figurado, explicações intuitivas representam as primeiras explicações em um diálogo explicativo que o falante tenta dar uma elucidação compreensível com base na falta de conhecimento do receptor.
Ponto(s) Forte(s)	Através de palavras-chave o método realiza buscas por conteúdos associados por representações semânticas na *Wikipedia*. Por exemplo, um profissional pode procurar informações relacionadas à "omoplata" no Código internacional de doenças (CID10) e obter informações pertinentes a anatomia do corpo humano. Isso pode facilitar a explicação do quadro clínico do paciente de forma intuitiva.
Limitações	As sentenças explicativas devem ser formalmente bem definidas para que o mecanismo de busca possa trazer conteúdos relevantes ou explicações intuitivas pertinentes ao contexto do diálogo. Os experimentos foram realizados com sentenças curtas como " *O dedo é parte do membro superior*". Embora a sentença seja de fácil compreensão, este acrônimo é parte do corpo humano e pode precisar de características adicionais para recuperar informações relevantes.
Referência	Forcher *et al.* (2014)
2- Título	***A Semantic Platform for Information Retrieval from E-Health Records***
Objetivo	Apresenta uma plataforma de recuperação de informação em PEPs através do uso de palavras-chave, ontologias e técnicas de busca semântica.

Ponto(s) Forte(s)	Os métodos propostos foram experimentados individualmente e logo após combinados. Os métodos envolvidos seguem: $MEDRUN_1$ (extração manual de palavras-chave), $MEDRUN_2$ (busca semântica utilizando conceitos UMLS), $MEDRUN_3$: (busca semântica utilizando conceitos ProMiner e CFR) e $MEDRUN_4$ (uso de ontologias). A abordagem proposta propicia a combinação estratégica dos resultados obtidos entre os experimentos.
Limitações	Embora os resultados dos experimentos tenham apresentado melhor desempenho para o método de recuperação proposto, compondo um número significativo de comparações entre os experimentos, as amostras (35) são relativamente pequenas para o contexto médico. Os resultados foram pouco significativos com relação a precisão considerando 0,5503 para mecanismos de buscas tradicionais e 0,5767 para o mecanismo de busca proposto no artigo.
Referência	Gurulingappa *et al.* (2011)
3- Título	***Efficient Medical Information Retrieval in Encrypted Electronic Health Records***
Objetivo	Apresenta uma abordagem para recuperação de informação em PEPs contendo documentos clínicos compartilhados em forma criptografada. O método baseia-se na exploração de meta-dados na descrição de documentos em conjunto com terminologias baseadas em padrão de linguagem médica incluindo *UMLS, SNOMED CT ou LOINC.*
Ponto(s) Forte(s)	O mecanismo de recuperação explora meta-dados adicionais (meta-informação) que analisa propriedades do *UMLS* antes de codificar o conteúdo. Promove assim a extração de conteúdos que revelam uma visão geral de conteúdos importantes para consultas, resultados de exames e a própria identidade do paciente.
Limitações	A abordagem apresentada foi avaliada em um conjunto pequeno de dados como prova de conceitos. Medidas de precisão e cobertura não foram levadas em consideração.
Referência	Pruski & Wisniewski (2012)
4- Título	***Using Semantic Search to Reduce Cognitive Load in an Electronic Health Record***
Objetivo	Apresenta através de experimentos a eficácia de um mecanismo de busca semântica em comparação a um mecanismo de busca tradicional em PEPs.
Ponto(s) Forte(s)	A busca semântica fornece sugestões dinâmicas de como o indivíduo realiza a consulta. A técnica de busca proposta explora sinônimos de termos médicos (*e.g.*, "ataque cardíaco" e "parada cardíaca"). Os resultados do experimento revelaram diferenças significativas aos métodos tradicionais. Os usuários foram capazes de navegar de forma mais precisa em termos de tempo (140 segundos de navegação na busca semântica versus 239 segundos de navegação na busca sintática) e número de cliques (11 cliques na busca semântica versus 35 cliques na busca sintática).
Limitações	O método proposto não considera a informação do histórico clínico do paciente, apenas representações semânticas sobre os termos da consulta realizada pelo médico. Nesse sentido, a percepção dos médicos relata problemas como "Prescrever uma medicação que o paciente seja alérgico". Este tipo de relevância do conteúdo não é levado em consideração pelo método proposto. Adicionalmente, o profissional tem que procurar o conteúdo digital e o prontuário em papel, resultando em dois

	processos.
Referência	Tawfik *et al.* (2011)
5- Título	***An Ontological Fuzzy Smith-Waterman with Applications to Patient Retrieval in Electronic Medical Records***
Objetivo	Define um sistema de apoio na recuperação de informação em PEPs. Explora técnicas de Inteligência Computacional como regras *fuzzy* em bases ontológicas para melhorar tomadas de decisões médicas.
Ponto(s) Forte(s)	Um experimento conduzido com 107 pacientes apresenta resultados significativos quando o método de busca consiste em recuperar tratamentos similares ou termos sinônimos considerando as classes ontológicas. Sob tal ótica, caso exista mais de um paciente com a mesma diagnose, o usuário pode consultar tratamentos realizados com sucesso ou tratamentos que não obtiveram êxito.
Limitações	A Taxonomia CID-9[22] explorada no método possui apenas 5 níveis de profundidade e os fatos sobre doenças são agrupadas em 56 códigos. Estes agrupamentos deixam de considerar diversos aspectos se comparados com o CID10 que possui atualmente 12450 códigos. Isso limita a granularidade e eficácia dos resultados recuperados. Outro fator que não foi considerado está relacionado aos sintomas do paciente para identificação de doenças e tomadas de decisão.
Referência	Popescu (2010)
6 -Título	***Ontology driven semantic profiling and retrieval in medical information systems***
Objetivo	Apresenta uma técnica para criação de subdomínios ontológicos (por especialidade médica). Isso através da extração de dados de um domínio (UMLS). Neste sentido, a criação dos perfis (subdomínios) propõe melhorar a validade semântica das informações requisitadas por profissionais de saúde.
Ponto(s) Forte(s)	A abordagem de filtragem de dados oferece a possibilidade de criação de um perfil avançado para usuários de um determinado domínio. Embora o trabalho seja criado para a área médica, ele não se limita somente a este domínio.
Limitações	A criação dos subdomínios ontológicos para cada especialidade médica (criação de perfil), possui restrições na abrangência do contexto em estudo pois restringi as informações a uma especialidade médica. Por exemplo, especifica pode deixar de relatar características de tratamentos eficazes realizados por outros profissionais. Experimentos com dados médicos computar medidas de precisão e cobertura não foram realizados.
Referência	Bhatt, Rahayu & Soni (2009)
7- Título	***CDAPubMed: a browser extension to retrieve EHR-based biomedical literature***
Objetivo	Desenvolve uma ferramenta objetivando facilitar a construção de consultas para recuperar a literatura científica relacionada com PEPs.

[22] http://tabnet.datasus.gov.br/cgi/sih/mxcid9lb.htm

Ponto(s) Forte(s)	A integração de palavras-chave contidas em prontuários médicos eletrônicos com informações da *Web* (HL7-CDA[23]) promove melhor indexação nos documentos pesquisados. Experimentos demonstraram que, por exemplo, em uma comparação com mais de 200.000 citações recuperadas por "neoplasia de mama", menos de dez citações foram recuperadas quando dez sintomas da doença dos pacientes foram adicionados usando CDAPubMed.
Limitações	Embora as medidas de recuperação propiciem uma melhor filtragem dos conteúdos, a relevância das informações retornadas pelo CDAPubMed não foi levada em consideração. No caso dos resultados dos experimentos, a abordagem não incorpora dados de outras fontes, *e.g.*, *openEHR*.
Referência	Rey *et al.* (2012)
8- Título	***Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers***
Objetivo(s)	Desenvolve modelos de representação de conhecimento combinando ontologia através da especificação de conjuntos de dados distintos. Visam apoiar a tomada de decisão em casos de cuidados integrados a pacientes com doenças crônicas.
Ponto(s) Forte(s)	Conceitos mapeados entre ontologias permitem que diversos termos (palavras-chave) associados a uma doença possam ser recuperados. O método proposto pode identificar, *e.g*, que para um quadro de "Diabetes", mais de 300 termos podem ser associados a este diagnóstico. Este método de recuperação pode, por exemplo, permitir ao médico, encontrar hipóteses de diagnósticos com base nos sintomas do paciente.
Limitações	O método proposto limita sua precisão em pesquisas de doenças mais específicas como "Diabetes mellitus tipo 2" (DM2). Em um contexto mais abrangente (variação de doenças existentes em PEPs), este fato pode dificultar as consultas no mecanismo de recuperação de informação.
Referência	Liaw *et al.* (2014)
9- Título	***Exploring the effectiveness of Medical Entity Recognition for Clinical Information Retrieval***
Objetivo(s)	Propõe um método para alavancar a recuperação de informação na área médica via extração de entidades do domínio em documentos não estruturados.
Ponto(s) Forte(s)	Explora o reconhecimento automático de entidades do domínio para estruturar textos livres da área médica e consultas nas bases. Para esse fim, investiga o uso de técnicas de aprendizagem de máquina na detecção de conceitos em consultas em língua natural que são transformadas em um formato estruturado.
Limitações	A proposta apresenta desempenho limitado (medidas de precisão e cobertura) quando os termos da consulta devem corresponder exatamente aos termos contidos nos documentos. Isso limita consulta de doenças em que os usuários tentam recupera-las através de termos relacionados.

[23] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

Referência	Cogley *et al.* (2013)
10- Título	***iHANDs: Intelligent Health Advising and Decision-Support Agent***
Objetivos	Proposta que explora o conceito de agentes inteligentes visando apoiar as tomadas de decisões em PEPs. O algoritmo definido realiza buscas em uma base de dados local e intercala os resultados com várias fontes de consultas na *Web*.
Ponto(s) Forte(s)	Através de uma pesquisa interativa as integrações entre a base local e as diferentes bases de consultas na *Web* propiciam ao usuário consultar doenças através dos sintomas apresentados pelos pacientes. Neste caso, a amostragem resultante das consultas fornecem informações de diversas bases distintas, incluindo-as em um repositório de dados que fornece dados estatísticos sobre a ocorrência de determinadas doenças.
Limitações	O número de estudos de casos foi limitado (4 pacientes). Os resultados das avaliações foram validados apenas por um profissional. As avaliações também não computaram as ações imediatas (procedimentos de urgência e emergência), que são as tomadas de decisões realizadas pelos profissionais no momento da consulta médica, *e.g.*, avaliar e medicar diante dos sintomas imediatos (não tratamentos posteriores).
Referência	Hannan *et al.* (2014)

As pesquisas analisadas apresentam uma visão ampla de técnicas de recuperação de informação em PEPs. O objetivo final das propostas é atender às necessidades dos usuários no momento da busca. Os trabalhos selecionados vão desde representações semânticas que promovem explicações intuitivas para profissionais da área médica (Forcher *et al.*, 2014), até métodos que otimizam a indexação de anotações em PEPs através de vetores que (criam mensagens de alerta) detectando tratamentos que necessitam ser acompanhados em uma unidade de tempo e que não foram concluídas, *e.g.*, Tratamento e acompanhamento em consulta "Prenatal[24] e Puerperal[25]".

Em uma análise mais refinada, nota-se que diversos trabalhos exploram a busca de conteúdos na *Web* como forma de enriquecer os termos das consultas (Bhatt, Rahayu & Soni, 2009). Por exemplo, resultados de busca retornam apenas descrições da doença com base na palavra-chave inserida (doença) do paciente através de conteúdos provindos da *Web* (Bo & Yang-Mei, 2015).

[24] Assistência na área da enfermagem e da medicina prestada à gestante durante os nove meses de gravidez.

[25] Consulta do puerpério é realizada entre a 4ª / 6ª semana após o parto.

Por outro lado, a recuperação de informação é pouco explorada em conteúdos descritos em linguagem natural que constituem a fonte essencial de informações em PEPs, e é o local no qual as intenções são expressas.

3.4 Discussão e Posicionamento

A análise aprofundada da literatura permitiu identificar que grande parte dos trabalhos se beneficia do uso de ontologias para representação semântica dos termos digitados nas consultas de buscas. De maneira geral, esses trabalhos ainda são protótipos que procuram desenvolver técnicas de recuperação de informação a partir de conteúdos padronizados na *Web* (*e.g.*, HL7, *OpenEHR*). Este fato limita as buscas por palavras-chave e sua representação semântica.

Apesar dos avanços obtidos pelos trabalhos pesquisados, sejam eles voltados à recuperação de informação utilizando intenções ou recuperação de informação em PEPs, analisamos que eles se limitam a ordenar os resultados em função da comparação das similaridades entre suas representações semânticas.

Alguns trabalhos objetivam capturar a intenção do usuário através de palavras-chave no momento da busca (Liaw *et al.*, 2014), mas não consideraram a existência de termos polissêmicos e sinônimos, o que pode levar ao retorno de resultados menos precisos. Os trabalhos raramente tratam as intenções de quem produziu o documento.

Na área da saúde, muitas avaliações desenvolvidas dos trabalhos não foram analisadas por médicos que possuem conhecimento no domínio para averiguar a relevância dos resultados de busca. Por outro lado, as investigações abordam aspectos relevantes a serem considerados na elaboração desta dissertação. Por exemplo, permitir que o profissional possa explicitar suas intenções no momento da consulta de busca e não apenas consultas compostas por palavras-chave. Adicionalmente, detectar a intenção do usuário que produziu a informação e a relação entre seus conceitos.

Esta análise aponta a originalidade desta dissertação que investiga e desenvolve técnicas de recuperação de informação utilizando fundamentos da Semiótica, modelos de representação de conhecimento e Teoria dos Atos de Fala. Nosso estudo explorará uma base de dados real em que os médicos irão tornar explícitas suas intenções no momento da definição das consultas de buscas.

No procedimento de recuperação, os elementos da consulta são comparados com anotações semânticas e de intenções, consideradas como parâmetros para a seleção e a ordenação dos resultados de busca.

3.5 Síntese do Capítulo

Este capítulo apresentou uma revisão da literatura elucidando o estado da arte nos temas em estudo desta dissertação. Foram apresentados trabalhos que visam estudar a recuperação de informação com base em intenção e detalhamos trabalhos que propõem técnicas de recuperação de informação em prontuários médicos eletrônicos.

Concluímos que há espaço de pesquisa para investigar mais profundamente a recuperação de informação em PEPs. Não foram encontrados trabalhos que lidam com elementos relacionados com intenção em máquinas de busca nesse contexto de aplicação. O próximo capítulo descreve a proposta de solução definindo os conceitos e técnicas desenvolvidos nesta dissertação.

Capítulo 4

Intenções na Recuperação de Informação: Análise, Métodos e Algoritmo

Este capítulo apresenta a proposta conceitual e o algoritmo originalmente desenvolvidos nesta dissertação como forma de aprimorar a recuperação de informação de registros médicos por meio do uso de classes de intenção. Com o objetivo de apresentar os fundamentos dos métodos e algoritmo desenvolvidos, a Seção 4.1 relata um estudo empírico sobre a análise das classes de ilocução em conteúdo de PEPs.

Esse estudo inicial foi essencial para o entendimento de como as intenções se manifestam em documentos médicos e a maneira que elas podem ser exploradas para o fim de recuperação de informação. A Seção 4.2 descreve a proposta de solução definindo do ponto de vista conceitual os elementos envolvidos no método de recuperação de informação. Como parte chave na solução proposta, a Seção 4.3 define o algoritmo de recuperação, seleção e ordenação proposto e sua execução é exemplificada na Seção 4.4.

4.1. Estudo sobre Intenções em PEPs

Este trabalho envolveu uma série de análises realizadas manualmente sobre um conjunto de PEPs reais para investigar como termos do domínio são usados por profissionais de saúde para expressar ilocuções relacionadas com as classes de intenção. O objetivo foi examinar se existem termos do domínio que são usados para expressar as diferentes dimensões utilizadas nas classificações das ilocuções conforme as teorias exploradas nesta investigação (*cf.*, Capítulo 2). Esta etapa foi útil para informar decisões nos elementos subsequentes deste trabalho.

Esta pesquisa considerou um conjunto de PEPs disponíveis em uma unidade de Pronto Atendimento de "Águas de Lindóia", no estado de São Paulo, Brasil. A quantidade total de PEPs do repositório foi de 13.300 onde todos os pacientes são anônimos. As análises manuais conduzidas utilizaram 26 casos relacionados ao diagnóstico da doença "Dengue [dengue clássico]". O estudo é composto por 5 análises

realizadas em sequência para revelar aspectos sobre as ilocuções nos textos médicos descritos nos campos de relato de pré-consulta e anamnese descritos em linguagem natural.

A análise 1 estuda os PEPs segundo a ocorrência de hipóteses diagnósticas para obter uma visão geral do conteúdo. Com base em um subconjunto de PEPs, a análise 2 efetua uma classificação manual das ilocuções para viabilizar os passos subsequentes. A análise 3 considera as ocorrências das dimensões do cubo de ilocuções, enquanto a análise 4 examina a frequência das classes de ilocução detectadas. Por fim, a análise 5 investiga termos específicos do domínio descobertos segundo os tipos de ilocução.

Análise 1. A pesquisa parte inicialmente da seleção de um conjunto de prontuários médicos que seja considerado relevante para examinar o papel das ilocuções. Foram selecionados subconjuntos dos PEPs de acordo com os diagnósticos (CID10) associados com cada PEP. A Tabela 4.1 apresenta a ocorrência dos 15 diagnósticos com maior incidência no conjunto total de PEPs. Apesar da distribuição de ocorrências nas categorias definidas, este estudo considerou em específico os diagnósticos de "Dengue [dengue clássico] --- (A90)", por possuir um bom índice de ocorrência (sexto mais frequente no conjunto analisado), por sua especificidade (as primeiras hipóteses, por exemplo, são demasiadamente genéricas) e por ser de fácil entendimento para os pesquisadores.

Tabela 4.1: Ocorrência de Hipóteses Diagnósticas nos PEPs

Hipótese Diagnóstica (CID10)	**Ocorrências**
Infecção aguda das vias aéreas superiores não especificada --- (J069)	526
Nasofaringite aguda [resfriado comum] --- (J00)	417
Dor lombar baixa --- (M545)	324
Náusea e vômitos --- (R11)	264
Laringofaringite aguda --- (J060)	191
Dengue [dengue clássico] --- (A90)	187
Tosse --- (R05)	174
Cefaléia --- (R51)	156
Sinusite aguda não especificada --- (J019)	148

Dor lombar baixa --- (M545)	133
Amigdalite estreptocócica --- (J030)	122
Infecção do trato urinário de localização não especificada --- (N390)	110
Alergia não especificada --- (T784)	110

Análise 2. Analisou manualmente os prontuários definindo as dimensões das ilocuções presentes nos documentos médicos. Os pesquisadores com o apoio de um médico envolvido na pesquisa examinaram o conteúdo dos registros para atribuir as dimensões de ilocuções nas sentenças do conteúdo. A Tabela 4.2 apresenta: (1) um exemplo de conteúdo em um PEP; e (2) a respectiva análise manual da descrição das ilocuções para três ilocuções, sendo as duas primeiras asserções e a última uma proposta.

Tabela 4.2: Classificação Manual das Ilocuções em PEPs

1	***EHR* - Dengue [dengue clássico] --- (A90)**			
	Paciente refere febre, cefaleia, dor retro ocular e mialgia há 1 dia. Nega alergia e comorbidade ao exame: Beg, corado, hidratado, eupneico, afebril restante sem alterações diagnóstico: Dengue? febre a/e pré-medicação: solicito hmg completo, paracetamol.			
2	**Classificação Manual das Ilocuções**			
	Trecho Analisado	**Dimensão**	**Valor**	**Classe**
I_1	*...REFERE* ***febre, cefaleia, dor retro ocular e mialgia*** *Há 1 DIA ...*	*Invenção (descritiva)*	Refere	**Asserção**
		Tempo (passado)	Há 1 dia	
		Modo (denotativo)	Refere	
I_2	*... NEGA* ***alergia e comorbidade*** *...*	*Invenção (descritiva)*	Nega	**Asserção**
		Tempo (presente)	Nega	
		Modo (denotativo)	Nega	
I_3	*... SOLICITO* ***hmg completo, paracetamol*** *...*	*Invenção (prescritiva)*	Solicito	**Proposta**
		Tempo (futuro)	Solicito	
		Modo (denotativo)	Solicito	

Análise 3. Com base nos dados gerados na análise 2, esta etapa envolveu um exame quantitativo sobre as ilocuções para identificar as ocorrências das dimensões. A Tabela 4.3 foi dividida em "*Zero*" e "*Um*". Segundo o referencial teórico adotado, consideramos "*Zero*" como passado/presente (dimensão tempo), descritivo (dimensão invenção) e denotativo (invenção modo), e "*Um*" como futuro (dimensão tempo), prescritivo (dimensão invenção) e afetivo (dimensão modo). Foram analisados manualmente 26 PEPs resultando em 201 ilocuções.

Tabela 4.3: Distribuição de ocorrência relativa ao Tempo, Invenção e Modo

	#Tempo	**#Invenção**	**#Modo**
Zero	182 (90.55%)	182 (90.5%)	186 (92.54%)
Um	19 (9.45%)	19 (9.45%)	15 (7.46%)

Resultados apontam que aproximadamente 90% das ilocuções estão no presente/passado, são descritivas e denotativas. O modo afetivo está presente em menos de 8% das mensagens.

Análise 4. A Tabela 4.4 apresenta a frequência de identificação dos tipos de ilocução nos textos analisados. As ilocuções identificadas nas mensagens médicas são 84,58% asserções, enquanto uma minoria consiste de propostas (7,96%), valoração (5,97%) e indução (1,49%). As classes de previsão, desejo, retratação e contrição não ocorreram no conteúdo dos documentos analisados.

Tabela 4.4: Frequência das Classes de Ilocução

Classe de Ilocução	**Frequência (Percentual)**
Asserção	170 (84.58%)
Proposta	16 (7.96%)
Valoração	12 (5.97%)
Indução	3 (1.49%)

Análise 5. Esta etapa examinou os termos representativos para cada classe de ilocução, com base no conteúdo dos prontuários. A Tabela 4.5 apresenta os principais termos detectados nas mensagens que indicam as ilocuções. Os termos "*Refere*", "*Apresenta*" e "*Relata*", por exemplo, estão presentes no total de 66 afirmações. Do ponto de vista qualitativo, esses termos foram utilizados tipicamente para confirmar sintomas ou doenças em pacientes.

Tabela 4. 5: Análise de termos expressos nas classes de ilocução

Termos	Asserção	Proposta	Valoração	Indução
Refere	48	-	3	-
Nega	82	-	2	-
Apresenta	12	-	-	-
Relata	6	-	-	-
Solicita	-	12	-	-
Recomenda	-	3	-	3
Há "*x*" dias/horas	24	-	2	-
Ausência	3	-	-	-
Queixa	-	-	2	-
Melhora	-	-	2	-
Outros	7	1	2	-

Os termos "Nega" e "Ausência" estão presentes no total de 85 afirmações e são usados frequentemente pelos médicos para relatar a ausência de sintomas ou doenças, que são importantes para o diagnóstico médico. Expressões relacionadas ao tempo (por exemplo, Há "*x"* dias/horas) estão presentes em 24 afirmações. As expressões sobre tempo são frequentemente usadas para referenciar a presença ou a falta de um sintoma ou doença em dias ou horas antes da consulta médica. Outros termos estão presentes em 7 afirmações (com uma ocorrência cada).

De maneira geral, este estudo apontou que a característica dos registros médicos é ser conciso e impessoal. Isto é observado pelo alto índice de ilocuções do tipo asserção. O aspecto mais relevante dos resultados alcançados é que há termos do domínio que melhor representam as dimensões das ilocuções. Por exemplo, "tosse há um dia" é um indicador diferente de "tosse há um mês". Essas observações são relevantes para a definição de consultas no processo de recuperação de informação. Necessitamos, portanto, possibilitar que a consulta permita definir um valor relacionado a uma dimensão como parâmetro de busca (por exemplo, tempo). As análises conduzidas

foram essenciais para tornar esse requisito explícito. O estudo empírico desenvolvido também é relevante para a definição de modelos que possam representar as ilocuções relacionadas com terminologias de domínio.

4.2. Método de Recuperação Informado por Ilocuções

A Figura 4.1 apresenta a visão geral da proposta de solução para um mecanismo de recuperação de informação que considere intenções. De maneira geral, os PEPs em um repositório (item D da Figura 4.1) são anotados com conceitos semânticos (item E da Figura 4.1) e tipos de ilocução (item C da Figura 4.1), expressando meta-dados adequados para o algoritmo de busca definido (item G da Figura 4.1).

No algoritmo de busca, além de conceitos advindos do domínio representados em Sistemas de Organização do Conhecimento (SOC) (item F da Figura 4.1), por exemplo o UMLS, o algoritmo tem como entrada uma consulta que define a necessidade do usuário considerando palavras-chave e tipos de ilocução (item A da Figura 4.1). Como resultado de saída, o algoritmo determina os registros relevantes ordenados (item H da Figura 4.1). Em seguida, detalhamos cada elemento da solução.

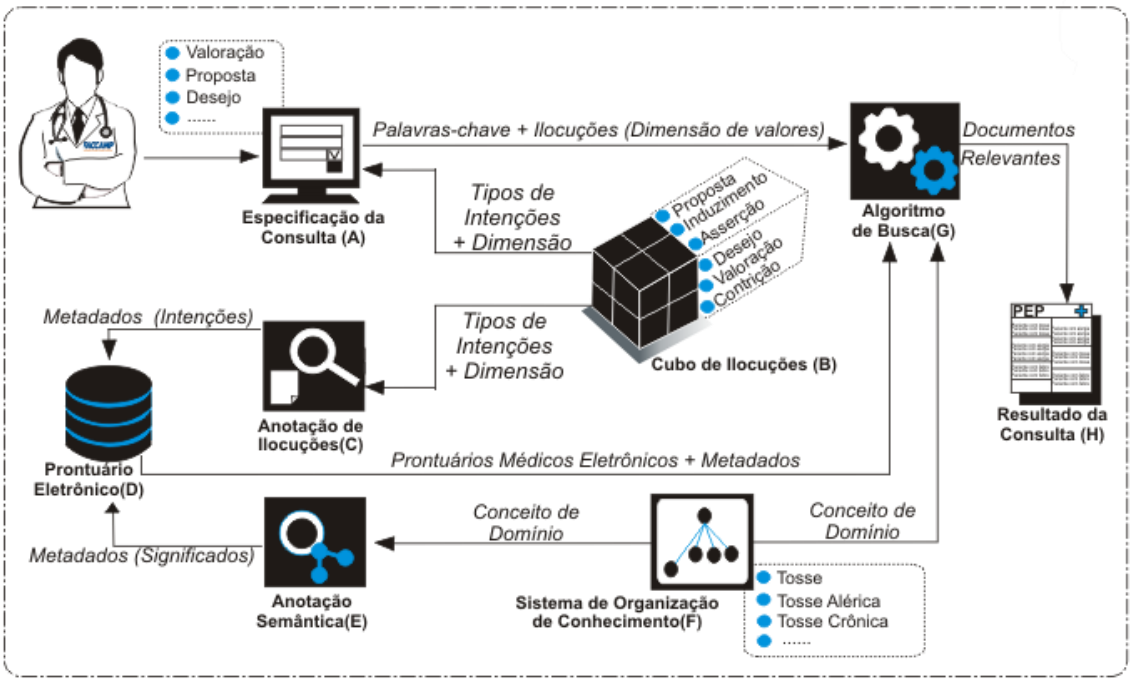

Figura 4.1: Visão Geral do Método de Recuperação de Informação

A Figura 4.2 detalha o processo de anotação semântica. Dado um conjunto de PEPs (item D da Figura 4.2), o método requer anotar explicitamente significados através

da detecção de conceitos em descrições textuais dos PEPs. A anotação semântica (item E da Figura 4.2) baseia-se na inspeção de conceitos definidos em SOCs para determinar os conceitos do domínio nos fragmentos de texto dos prontuários. Os SOCs incluem um conjunto de conceitos denotados por um código de identificação e rótulos.

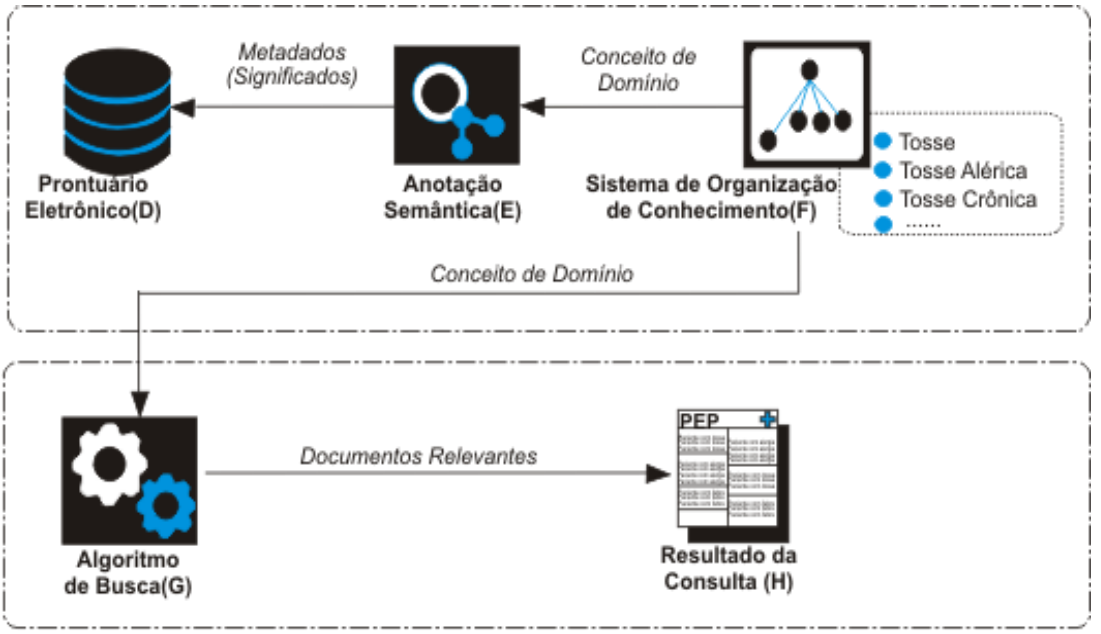

Figura 4.2: Processo de Anotação Semântica

Os conceitos são organizados através de relações, por exemplo, "é-um", "parte-todo", "relacionado-com". Ao relacionar fragmentos de texto com códigos de conceitos em SOCs, gera-se meta-dados semânticos para os prontuários. Por exemplo, no seguinte fragmento de texto *"Paciente em tratamento e avaliação da asma durante sete dias",* o termo "*asma*" é detectado como um conceito: Asma (doença) com código: "SCTID :195967001" na *SNOMED CT*.

O tratamento automático de anotações semânticas é estudado na literatura e há diversas ferramentas construídas para esse fim. Por exemplo, a ferramenta *MetaMap*[26], desenvolvida pela *NLM*, é utilizada para a anotação de textos da área médica explorando os SOCs que fazem parte do UMLS.

Este trabalho envolveu o reuso de soluções existentes para a tarefa de anotação semântica. No entanto, foi necessário lidar com desafios de registros descritos na Língua

[26] https://metamap.nlm.nih.gov/

Portuguesa. Diversas ferramentas que exploram processamento de linguagem natural para fins de anotações utilizam a estrutura da língua Inglesa como base. Em nossa solução, definimos um procedimento que efetua requisições a um conjunto de serviços e padrões de programação definidos para acesso ao UMLS. Exploramos os SOCs integrados ao UMLS que possuem versão em língua Portuguesa (*cf.* Seção 5.1.2).

A Figura 4.3 apresenta o processo de anotação de intenções. A anotação de ilocução (item C da Figura 4.3) consiste em detectar tipos de ilocução e as dimensões do cubo (item B da Figura 4.3) em fragmentos de texto descritos em linguagem natural dos prontuários eletrônicos. Por exemplo, nos fragmentos de texto "... *REFERE* ***FEBRE, CEFALEIA, DOR RETRO OCULAR E MIALGIA*** ..." e "... *NEGA* ***ALERIA E COMORBIDADE*** ..." (*cf.* Tabela 4.2) podem ser anotados como o tipo de ilocução "asserção", (tempo: presente / passado, invenção: descritivo e modo: denotativo). O fragmento de texto "... *SOLICITO* ***HMG COMPLETO, PARACETAMOL*** ..." pode ser anotado com o tipo de ilocução "proposta" (tempo: futuro, invenção: prescritiva e modo: denotativo). Em adição ao nível de sentenças, outras estruturas do documento podem ser anotadas como parágrafos, título e rótulos de ilocução atribuídos para o documento em geral.

Combinar os resultados da anotação semântica com a anotação de ilocuções é um ponto chave na originalidade desta pesquisa. Contudo, detectar de maneira totalmente automática e com acurácia no conteúdo dos PEPs os diferentes tipos de ilocução é um desafio de pesquisa em longo prazo. Para viabilizar esse componente da solução, nesta dissertação investigamos preliminarmente o potencial de termos específicos do domínio para expressar as dimensões de ilocuções (*cf.*, Seção 4.1).

O processo de anotação pode ser assistido por um modelo que represente a relação entre os termos do domínio que melhor caracterizam as dimensões do "cubo". Visando investigar experimentalmente a influência das anotações na abordagem de recuperação proposta, desenvolvemos funcionalidades no sistema de *software* implementado nesta dissertação que permite ao médico declarar as anotações de ilocução manualmente, conforme descrito no Capítulo 5.

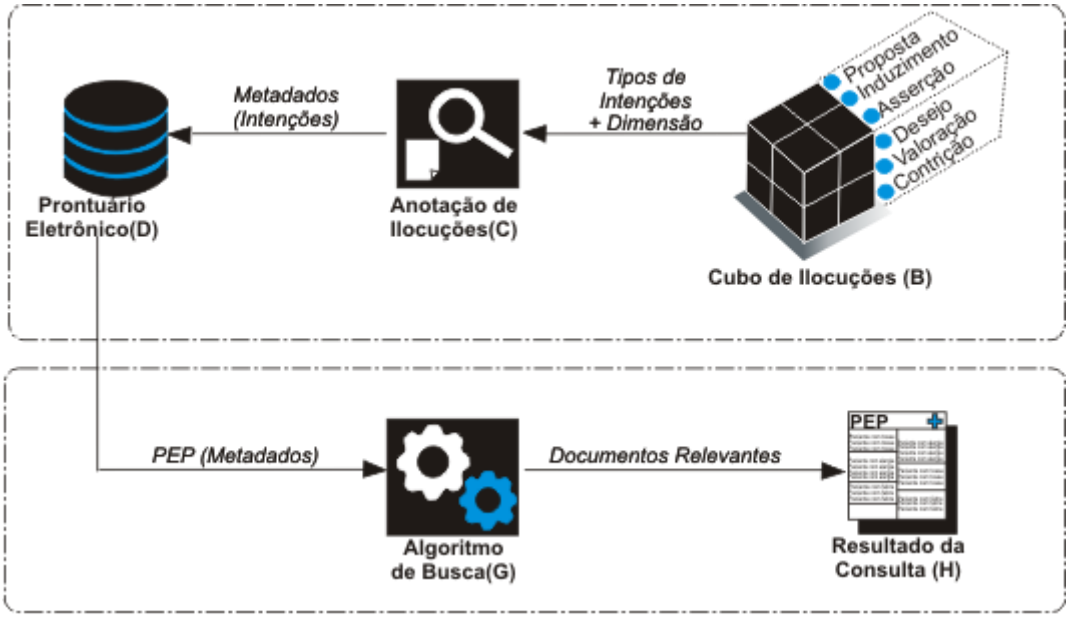

Figura 4.3: Processo de Anotação de ilocuções

A Figura 4.4 apresenta a especificação da consulta de busca. Além do conjunto de PEPs que deve ser recuperado ordenadamente, e os meta-dados de anotação, o algoritmo de busca requer a especificação de uma consulta como entrada para sua execução. Neste trabalho, a consulta envolve incluir os tipos de intenção e as dimensões definas pelo "cubo". De acordo com nossas análises experimentais de fundamento, esses elementos da consulta podem desenvolver um papel chave na recuperação informada por ilocuções.

Na solução definida, profissionais de saúde especificam a consulta, considerando: 1) um conjunto de palavras-chave relacionadas a doenças, sintomas ou qualquer outro aspecto que julgar relevante; 2) a dimensão da ilocução para uma determinada palavra-chave; e 3) os valores quantitativos relativos a esta ilocução. Por exemplo, palavra-chave: "*Tosse*" com a dimensão "*Tempo*" expressa por "*mais de quatro semanas*".

A seção seguinte descreve o algoritmo de busca (item G da Figura 4.4) que filtra e ordena um conjunto de PEPs relevantes a partir de toda a base disponível.

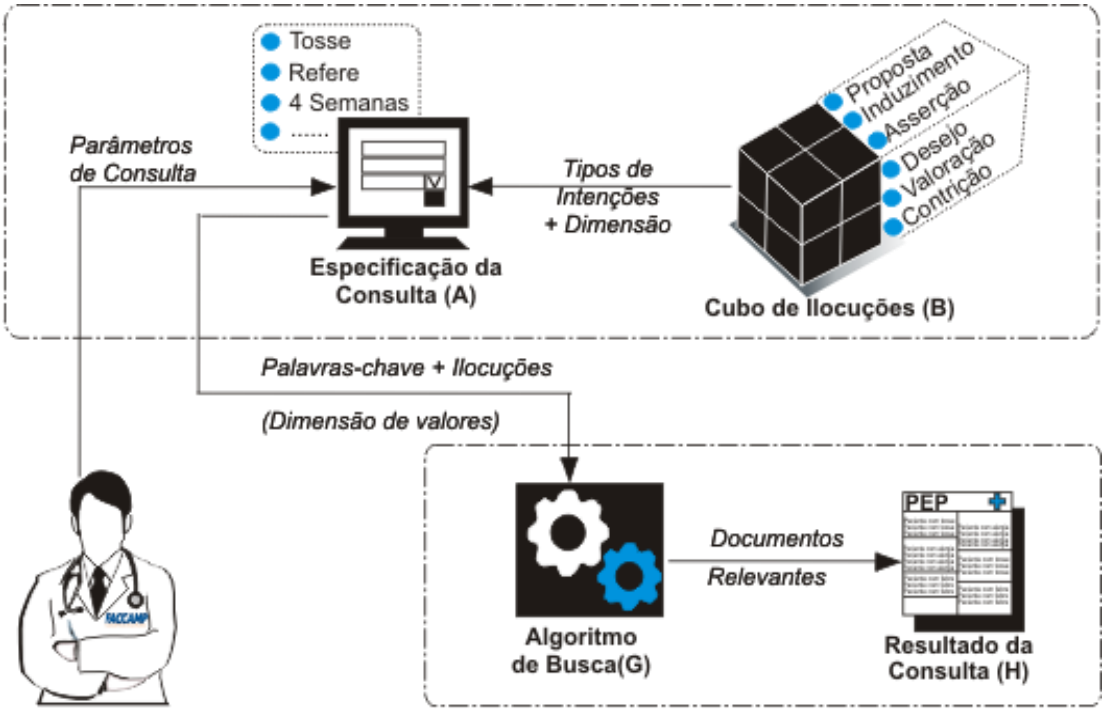

Figura 4.4: Especificação de Consultas de Busca

4.3. PraSA - ***Pragmatic Search Algorithm***

O algoritmo ***PraSA*** (*Pragmatic Search Algorithm*) proposto visa filtrar e ordenar um conjunto de PEPs. O algoritmo recebe como entrada os seguintes elementos: (1) o repositório de PEPs; (2) os meta-dados gerados pelas anotações; (3) SOCs do domínio; e (4) a consulta que denota palavras-chave, dimensão de ilocução e valores para as dimensões. Definimos, assim, sete etapas principais no algoritmo: (1) Recuperação Inicial; (2) Consulta de termos relacionados em SOCs; (3) Recuperação Expandida; (4) União de Resultados; (5) Filtragem com base em ilocução; (6) Ordenação; e, (7) Prioridade de Ocorrência.

Em seguida, essas etapas são detalhadas e apresentamos a formalização do algoritmo. A Figura 4.6 define o algoritmo; como entrada temos R como um conjunto de documentos $d_{i,\ i \in N}$, $R = \{d_1, d_2, \ldots, d_n\}$. O conjunto S define os conceitos c_i dos SOCs, sendo $S = \{c_1, c_2, \ldots, c_k\}$. Definimos a consulta como um conjunto de elementos, tal que, $C = \{E_1, E_2, \ldots E_x\}$ e E_i é uma tripla denotando palavra-chave (p), dimensão (dm) e valor (vl) da ilocução, assim, $E_i=(p_i, dm_i, vl_i)$. Os meta-dados são definidos pelo conjunto M que expressa as anotações. Logo, $M=\{A_1, A_2, \ldots, A_y\}$, sendo $A_i=(d_i, st_i, dm_i, vl_i)$. O

elemento d_i representa o documento, enquanto st_i uma sentença e os parâmetros restantes a dimensão (*dm*) e valor da ilocução (*vl*).

Passo 1 – Recuperação Inicial: Esta etapa recupera a partir do repositório de PEPs todos os documentos em que existam ocorrências de palavras-chave definidas na consulta do usuário por meio de busca léxico-sintática. Por exemplo, se a consulta define "tosse" e "alergia" como palavras-chave, essa etapa recupera todos os registros que contenham explicitamente esses termos incluindo declinações/flexões. O resultado de saída é representado pelo conjunto REL_{INI}. A ordenação básica de resultados nesta fase reflete os métodos implementados na plataforma de busca sintática reutilizada em nossa solução (o Capítulo 5 descreve o uso da plataforma no sistema). A linha 6 do algoritmo *PraSA* (Figura 4.6) representa a recuperação inicial para cada palavra-chave.

Passo 2 – Consulta de termos relacionados em SOCs: Para permitir uma busca expandida com termos relacionados ao domínio, o algoritmo verifica a ocorrência das palavras-chave em conceitos presentes em SOCs (Por exemplo, explorando o *UMLS*). Esta etapa detecta e seleciona vocabulários relacionados com as palavras-chave definidas. O algoritmo pode selecionar termos sinônimos e semanticamente relacionados com as palavras-chave iniciais. Por exemplo, para o termo "alergia" são recuperados termos como "atopia", subtipos de "alergia" e outros termos relacionados, ou seja, retorna termos de acordo com os relacionamentos descritos no SOC. No algoritmo PraSA (Figura 4.6), dentro do laço de repetição inicial, na linha 7, os termos são recuperados pela função *EXPANSAO*.

Passo 3 – Recuperação Expandida: Nesta etapa, o algoritmo efetua uma nova busca no repositório de PEPs selecionando registros que contenham ocorrência dos termos obtidos na consulta aos SOCs. Novamente, a busca, considerando os termos semânticos, é efetuada na plataforma de busca que indexa os PEPs. Neste instante, se a consulta é definida por diversas palavras-chave, diversos termos são obtidos para cada palavra-chave. Nesse caso, o algoritmo efetua as buscas através de uma combinação de todos os termos expandidos obtidos. Por exemplo, se para a palavra-chave inicial P_a foi encontrado os termos relacionados T_{Pa1} e T_{Pa2} e na mesma consulta, a palavra-chave inicial P_b possui os termos relacionados T_{Pb1} e T_{Pb2}, logo a recuperação expandida

consulta "T_{Pa1} *e* T_{Pb2}" *e* "T_{Pa2} *e* T_{Pb1}". Os resultados encontrados são representados pelo conjunto REL_{EXP} na linha 11 do algoritmo PraSA.

Passo 4 – União de Resultados: Esta etapa faz a união do conjunto REL_{INI} com REL_{EXP}. O resultado de saída é representado pelo REL_U, onde, primeiramente os elementos do conjunto REL_{INI} são dispostos em ordem anterior aos exclusivos aos elementos do conjunto REL_{EXP}. A união é efetuada na linha 13 do algoritmo PraSA.

Passo 5 – Filtragem com base em ilocução: Até o presente momento, apenas palavras-chave e termos relacionados foram usados para recuperar PEPs. Nesta fase, o algoritmo explora parâmetros adicionais da consulta para obter um conjunto de resultados mais significativos. Considerando que, uma consulta inicial envolve palavras-chave, dimensão e valor, o algoritmo expande esta tripla para todos os termos associados às palavras-chave (*cf.* Figura 4.5).

O <termo> significa uma ou mais palavras-chave (*keyphrase*), sinônimos ou termos relacionados terminologicamente, enquanto <dimensão> denota a dimensão do tipo de ilocução especificado na consulta. O parâmetro <valor> descreve o valor quantitativo ou expressões que representam a dimensão no domínio. Por exemplo, palavra-chave "*artralgia*", com dimensão "*tempo*" e valor "*dois dias*".

O algoritmo, na linha 15, filtra PEPs presentes em REL_U que tenham anotações de ilocuções na qual termos aparecem em fragmentos de texto dos registros anotados com a dimensão do tipo de ilocução e seu respectivo valor de acordo com os parâmetros definidos na consulta de entrada. O resultado filtrado é representado pelo conjunto REL_F. Para considerar as anotações, a função FILTRAR no algoritmo recebe como parâmetro os meta-dados *M*.

Figura 4.5: Parâmetros na filtragem de resultados de busca

Passo 6 – Ordenação: Se algum documento existe em REL_F, o algoritmo PraSA, na linha 17, ordena os PEPs de acordo com o tipo de ocorrências entre os termos e ilocuções encontradas na etapa de filtragem. Primeiro, o algoritmo prioriza os PEPs que contenham exatamente as palavras-chave da consulta de entrada encontradas na anotação da ilocução. Em segundo, o algoritmo ordena os PEPs, na qual termos relacionados e sinônimos do conjunto *T* são encontrados de acordo com a ilocução anotada (denominamos recuperação pragmática aproximada). Portanto, os documentos em que as instâncias de ilocuções ocorrem com palavras-chave, sinônimos ou termos conceitualmente relacionados obtém maior relevância na ordenação dos resultados.

Algoritmo 1: Busca Pragmática

```
    Entrada: R, S, M, C
    Saída: REL'_F
 1  Início
 2    REL_INI = vazio
 3    T = vazio
 4    para cada E ∈ C faça
 5      para cada p ∈ E faça
 6        REL_INI ← REL_INI ∪ BUSCA - SIN ( p, R )
 7        T ← T ∪ EXPANSAO ( p, S)
 8      fim
 9    fim
10    para cada t ∈ T faça
11      REL_EXP ← REL_EXP ∪ BUSCA - SIN_EXP ( t, R, T )
12    fim
13    REL_U = REL_INI ∪ REL_EXP
14    para cada E ∈ C faça
15      REL_F ← FILTRAR ( REL_U, E, T, M)
16    fim
17    REL'_F = ORDENAR ( REL_F, C, M)
18  fim
19  retorna REL'_F
```

R = Conjunto de Documentos $\{ d_1, d_2, ..., d_n \}$
S = Conjunto de Conceitos $\{ c_1, c_2, ..., c_k \}$ *tal que*
C = Conjunto de Elementos $\{E_1, E_2, ...E_x\}$ *denotando palavra-chave (p), dimensão (dm) e valor (vl)*
M = Conjunto de Metadados $\{ A_1, A_2, ..., A_y \}$
$A = \{d_i, st_i, vl_i\}$, *tal que* d_i *= um documento,* st_i *= uma sentença,* dm_i *uma dimensão e* Vl_i *= valor da ilocução*

Figura 4.6: Descrição do Algoritmo PraSA

Passo 7 – Prioridade de Ocorrência: O algoritmo melhora a ordenação de documentos em REL_F de acordo com o número de ocorrências de termos, independentemente da sua localização em relação às ilocuções anotadas. Isto significa que os documentos com maiores ocorrências de palavras-chave, ou sinônimos dentro e fora da anotação de ilocuções tem maior prioridade no algoritmo. Esta prioridade é determinada, por exemplo, via métodos de busca sintática, como na solução apresentada no próximo capítulo. O algoritmo PraSA gera o resultado final definido por REL'_F. O

resultado final do algoritmo consiste em um subconjunto relevante e ordenados dos PEPs do repositório.

4.4. Ilustrando a Execução do Algoritmo

Nesta investigação, escolhemos em parceria com profissionais de saúde um cenário típico para ilustrar a execução do algoritmo proposto. Exploramos a base real de PEPs de unidades básicas de saúde no município de Aguas de Lindóia, estado de São Paulo, Brasil, conforme descrito anteriormente. O objetivo do cenário definido foi "*Recuperar todos os registos de pacientes que relataram tosse por mais de quatro semanas e negam qualquer tipo de alergia*". A Tabela 4.6 apresenta exemplos de PEPs relacionados com o cenário proposto.

Tabela 4.6: Exemplos de PEPs para o cenário proposto

Prontuário	Anamnese	Recuperar
14569	Mãe refere quadro de tosse produtiva a cerca de 2 semanas, com piora há 2 dias. Refere febre não aferida, não soube informar quando. Fez uso de ibuprofeno, com melhora do quadro. Refere alergia medicamentosa de amoxicilina. Ao exame: beg, eupneica, afebril, hiperemia de orofaringe, mv presente e simétrico, sem ruídos adventícios. FR: 29 ipm. pre-medicacao: oriento procurar acompanhamento em ubs.	Não
8440	Tosse há mais de 4 semanas; ap: nega alergia medicamentosa; nega pat prévias, exceto hiperuricemia, em uso de alopurinol 300 mg/dia; ao exame: - ar: mv + bilat sem ra.	Sim
11934	Paciente relata há mais de 4 semanas quadro gripal: dores em rosto, dores em orofaringe, tosse seca com piora noturna, nega febre, nega atopia.	Sim
13052	Paciente refere tosse seca, dispneia, e coriza amarelada ha mais de 4 semanas. nega febre e demais queixas. nega alergia ao exame. Diagnostico: sinusite? tosse alergica?. Pre-medicacao: amoxicilina 500mg 8/8h durante 7 dias acebrofilina xarope 12/12h 7 dias oriento retorno caso não haja melhora do quadro ou caso comece apresentar qualquer sinal ou sintoma de pneumonia.	Sim

O cenário possui uma motivação médica explícita, pois a presença de tosse por mais de quatro semanas, sem que o paciente apresente qualquer sintoma de alergia, direcionam o caso para a investigação mais aprofundada sobre possível tuberculose. Nesse caso, devem ser observadas em situações específicas e tratadas com maior prioridade. Um médico auditor pode, por exemplo, investigar a ocorrência dos sintomas em uma determinada população ou um pesquisador poderia relacionar os dados com outras comorbidades.

Conforme descrito, as consultas no mecanismo de recuperação desenvolvido são formadas por três parâmetros de entrada, compostos por uma palavra-chave (por exemplo, "tosse" e "alergia"), uma dimensão da ilocução que a palavra-chave está relacionada (ou seja, tempo, invenção ou modo) e um valor que específica à dimensão (por exemplo, para tempo, "quatro semanas", e para o modo, "nega"). A consulta no cenário foi construída usando a seguinte notação:

1- **Tm** < Tosse> + **Dm** <Tempo> + **Vl** <Mais de 4 semanas>

2- **Tm** <Alergia> + **Dm** <Modo> + **Vl** <Nega>

Na execução do algoritmo, temos o resultado das seguintes etapas:

1. A recuperação inicial retorna todos os registros contendo as palavras-chave – "Tosse" e "Alergia" utilizando busca sintática. Nesta etapa, o algoritmo não recupera o prontuário 11934, pois não contem os termos "Tosse" e "Alergia", já os demais PEPs da Tabela 4.6 são recuperados pois possuem esses termos.

2. Os termos "Tosse" e "Alergia" são pesquisados, respectivamente, nos SOCs existentes (por exemplo, no UMLS "Tosse" tem o CUI[27] "CO0010200" e "Alergia" possui CUI "C1527304"). Sinônimos e termos relacionados são recuperados a partir do conteúdo dos SOCs;

3. A busca expandida recupera no repositório de PEPs todos os documentos contendo os sinônimos e termos relacionados obtidos das palavras-chave "*Tosse*" e "*Alergia*" (por exemplo, "*hipersensibilidade*" e "*atopia*"). Neste caso, o prontuário 11934 é recuperado, pois possui conceitos relacionados às palavras-chave de entrada, *e.g.* "Atopia" que é um conceito de alergia.

4. O algoritmo realiza a união dos resultados da busca inicial ("*Tosse"* e "*Alergia*") com os resultados da busca expandida. Os PEPs da recuperados inicialmente na primeira etapa são inseridos em ordem antes dos PEPs da busca expandida. Partindo deste princípio, o algoritmo ordena a seguinte sequência para as ocorrências encontradas na Tabela 4.6:

[27] Código único de identificação de um conceito que representa o significado no domínio.

• Primeiramente o algoritmo seleciona o PEP (8840) como prioridade por possuir os termos relacionados às palavras-chave “Tosse” e “Alergia”.

• Em seguida o algoritmo seleciona os PEPs que possuem os termos relacionados às palavras-chave da entrada da consulta *e.g.*, “Tosse Produtiva” e “Tosse Seca” e “Atopia”.

5. Os registos presentes no conjunto resultante da união são filtrados de acordo com as anotações de ilocuções previamente cadastradas no repositório (O Capítulo 5 detalha o processo de anotação das ilocuções). O algoritmo considera apenas os registros com o termo “Tosse” (incluindo sinônimos e os termos relacionados com “Tosse”) anotados com a dimensão “Tempo”, relacionados com “mais de quatro semanas” e o termo “Alergia” (incluindo sinônimos e os termos relacionados de “Alergia”) anotado com a dimensão “Modo”, relacionada com o valor “Nega”.

Os registros com anotações contraditórias são removidos a partir dos resultados. Por exemplo, “*Alergia*” anotado com dimensão “Modo”, relacionada com “Refere” (ou outro termo que indica a presença de alergia em vez de sua falta) e dimensão “Tempo” relacionada com “menos de quatro semanas”;

6. Os registos médicos são ordenados de acordo com as palavras-chave e ilocuções. Documentos que incluem as palavras-chave “*Tosse*” e “*Alergia*” anotadas respetivamente nas ilocuções são colocados em primeiro lugar. Posteriormente, os sinônimos de “*Tosse*” e “*Alergia*” com anotações correspondentes estão incluídas nos resultados. Eles são seguidos por termos relacionados com anotações correspondentes.

O seguinte trecho ilustra o resultado de um texto recuperado de um caso relacionado ao histórico do paciente no cenário proposto:

“*Paciente relata há mais de 4 semanas quadro gripal: dores em rosto, dores em orofaringe, tosse seca com piora noturna, nega febre, nega atopia.*”.

Nesse exemplo, um algoritmo que apenas se fundamenta em uma recuperação sintática e procura por “Tosse” e “Alergia” não pode recuperar o registro. Na busca expandida, o termo “Atopia” é recuperado pela expansão de consulta como termo relacionado. Contudo, ainda não é capaz de efetuar a relação do termo “Tosse” com o tempo e da negação da “Alergia”. A anotação de ilocução confirma a existência de

“Tosse” “há mais de 4 semanas”, e a “ausência” de sintomas relacionados à “alergia”. Esse registro é ordenado logo após os resultados de documentos com os termos exatos “*Tosse*” e “*Alergia*”, se existirem.

O texto a seguir ilustra um registo que não é recuperado pelo algoritmo proposto:

“*Mãe refere quadro de tosse produtiva a cerca de 2 semanas, com piora há 2 dias. Refere febre não aferida, não soube informar quando. Fez uso de ibuprofeno, com melhora do quadro. Refere alergia medicamentosa de amoxicilina*.”

Neste caso, um algoritmo apenas explorando a busca inicial para “*Tosse*” e “*Alergia*” pode retorna tal registo. Contudo, nosso algoritmo analisa as anotações de ilocução dos registros e permite detectar a “presença” de “alergia” e a “ausência” de “tosse” durante mais de quatro semanas. Isto contradiz os parâmetros de consulta e, consequentemente o PEP 14569 (*cf.* Tabela 4.6) é removido do conjunto de resultados.

4.5. Síntese do Capítulo

Este capítulo primeiramente apresentou análises experimentais que fundamentaram a proposta de solução de um mecanismo de recuperação que considera aspectos relacionados a intenções do usuário. Por meio dessas análises podemos entender diversas maneiras de como as ilocuções se apresentam em registros médicos.

As análises revelaram a relevância dos termos que indicam as dimensões de uma ilocução. Por exemplo, uma afirmação que relata a existência de uma doença é completamente diferente do ponto de vista médico de uma afirmação que nega a mesma doença. A modelagem e interpretação computacional desses aspectos podem resultar em melhorias de ferramentas de recuperação de informação médica e sugeriram elementos relevantes na solução definida nesta pesquisa.

Este capítulo definiu um método de recuperação de informação que combina o significado de termos com tipos de intenção. Os resultados obtidos permitem especificar termos relacionados do domínio com dimensões de ilocuções nas consultas especificadas. Isso gera novas possibilidades de recuperação considerando a intenção dos usuários na consulta por PEPs. Definimos e formalizamos um algoritmo de busca que possibilita filtrar e ordenar os resultados de busca. Apresentamos um cenário que

exemplifica a execução do algoritmo e realça as características do seu comportamento ilustrando casos de registros recuperados e não recuperados.

Embora os resultados alcançados sejam relevantes, esta pesquisa pode avançar em diversas direções. Por exemplo, investigar um modelo que represente aspectos sobre tempo e o relacionar com a dimensão tempo. É igualmente relevante a pesquisa de um mecanismo automático de anotação de ilocuções que deve requerer lidar com diversos desafios em processamento de linguagem natural. O próximo capítulo apresenta aspectos de implementação da solução definida em um sistema de *software*.

Capítulo 5

SiRBI: O Sistema de Recuperação com Base em Intenções

Neste capítulo apresentamos o **Si**stema de **R**ecuperação com **B**ase em **I**ntenções (SiRBI). Esse sistema tem o objetivo de recuperar PEPs considerando elementos de intenção expressos nas consultas dos usuários. Definimos a arquitetura do sistema e descrevemos os componentes utilizados no desenvolvimento. A Seção 5.1 detalha a arquitetura, os componentes e as tecnologias envolvidas na implementação do SiRBI. A Seção 5.2 reporta as funcionalidades junto com as principais interfaces de interação com o usuário. Finalmente, a seção 5.3 apresenta uma síntese das contribuições do capítulo.

5.1. Arquitetura, Componentes e Tecnologias Empregadas

A arquitetura definida para o sistema inclui diversos componentes e plataformas. Primeiramente, esta seção apresenta uma visão geral da arquitetura, detalhando a relação entre os componentes. Em seguida, serão apresentados detalhes envolvendo técnicas e tecnologias de cada componente de maneira individual.

A Figura 5.1 apresenta uma visão geral da arquitetura. O componente de análise de termos (item A da Figura 5.1) processa os PEPs descritos em linguagem natural de modo a efetuar a separação dos termos chave dos documentos. Isso é relevante para indexação e apoio ao procedimento de expansão de consultas. Dado um conjunto de PEPs como entrada, o componente separa cada termo dentro do conjunto de PEPs considerando como, padrão para a separação, os espaços contidos entre cada termo. Como resultado, o componente de análise de termos armazena os termos em uma base de dados (Item B da Figura 5.1) descartando *stopwords* (palavras consideradas "vazias", *i.e.*, que não fazem parte do vocabulário médico relevante ao contexto de PEPs).

O componente de expansão de consultas (item C da Figura 5.1) visa detectar termos semanticamente relacionados àqueles identificados no conteúdo dos PEPs. Para este fim, o componente seleciona os termos processados pelo componente de análise de termos (item A da Figura 5.1) e busca os códigos em um SOC como uma ontologia (item D da Figura 5.1). Neste trabalho, exploramos o UMLS para auxiliar a expansão

semântica de consultas. Para cada termo, um conjunto de CUIs relacionados são recuperados e armazenados no banco de dados (item B da Figura 5.1).

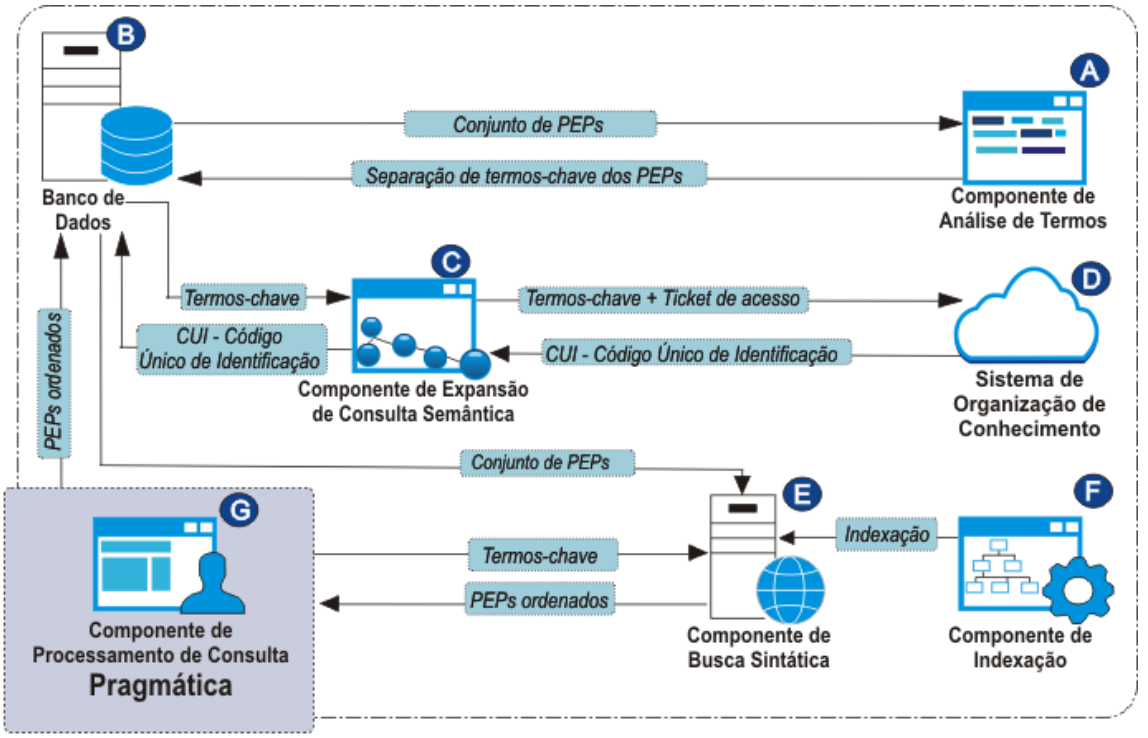

Figura 5.1: Visão geral da arquitetura do SiRBI

O conjunto de PEPs é indexado pelo componente de indexação (item F da Figura 5.1) para posterior uso pelo processo de recuperação. Os componentes definidos nos itens A, C e F da Figura 5.1 são executados em um momento anterior as consultas, com o objetivo de indexar e preparar os dados para acomodar as consultas. O componente de principal contribuição deste trabalho é definido pelo item G na Figura 5.1. O processamento de busca pragmática explora o componente de busca sintática (item E da Figura 5.1) para a consulta do índice e ordenação primária dos resultados da consulta. Nas seções subsequentes, detalhamos o funcionamento de cada componente da arquitetura proposta.

5.1.1. Análise de Termos

A Figura 5.2 detalha o componente de análise de termos (item A da Figura 5.1), que tem como objetivo processar a separação dos termos chave no conjunto de PEPs. O

módulo de análise de termo (item A da Figura 5.2)[28] seleciona os PEPs existentes no banco de dados (item C da Figura 5.2). Para cada PEP, em um laço de repetição (item B da Figura 5.2), o conteúdo descrito em linguagem natural do PEP é processado sentença a sentença. Os termos-chave são processados para cada PEP. Por exemplo, em uma sentença "*Paciente refere febre, vômitos e dor*" identificada pelo algoritmo, o módulo (item D Figura 5.2) processa a sentença e tem como saída os seguintes elementos: "**Paciente**", "**refere,** "**febre**", "**,**" , "**vômitos**", "**e**", "**dor**".

Os termos separados são enviados para o procedimento (*stored procedure*) de análise de termos (item E da Figura 5.2). Este procedimento é responsável por processar os termos e eliminar as palavras vazias (*stopwords*) das sentenças. Um conjunto inicial de palavras vazias presentes na ferramenta *Lucene Apache Solr*[29] foi considerado, posteriormente outras palavras foram inseridas considerando o domínio de saúde. Os termos não descartados são armazenados na base de dados. No exemplo apresentado na Figura 5.2 os elementos "Paciente", "refere", "," e "e" foram classificados como palavras vazias e, portanto, não foram incluídos na base.

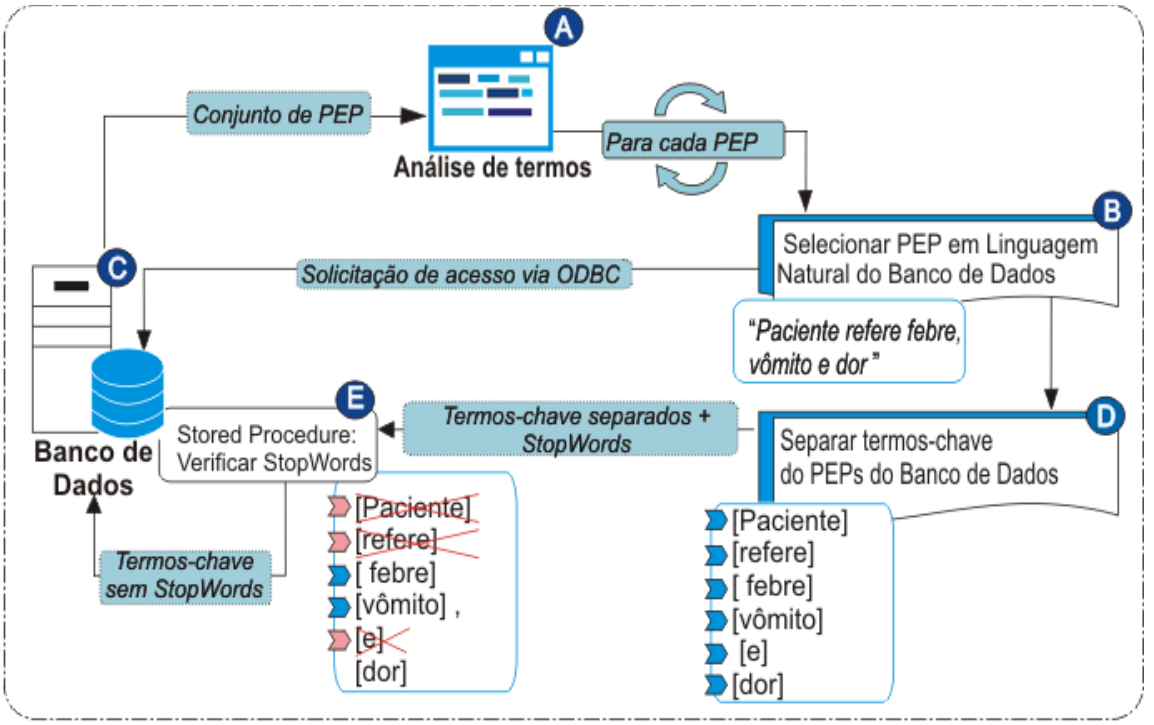

Figura 5.2: Componente de análise de termos

[28] Módulo construído utilizando a plataforma de desenvolvimento Delphi
[29] http://lucene.apache.org/solr/

5.1.2. Expansão de Consultas

A Figura 5.3 descreve o componente de expansão de consulta (item C da Figura 5.1), que objetiva identificar termos semanticamente relacionados aos termos identificados no conteúdo dos PEPs. Para este fim, o módulo de expansão de consulta (item A Figura 5.3) explora uma API[30] (*Application Programming Interface*) que permite a manipulação e o acesso de aplicativos computacionais a um Sistema de Organização do Conhecimento. Em particular, este trabalho explora a conexão via serviços Web ao servidor UTS[31] (*UMLS Terminology Services*).

O sistema seleciona os termos resultantes do processo de análise de termos[32] realizado pelo componente descrito anteriormente e usa a API para acesso remoto junto ao servidor de terminologia. Para cada termo, a API gera um código (*ticket* de acesso) para validar a entrada deste termo-chave do PEP junto ao UTS.

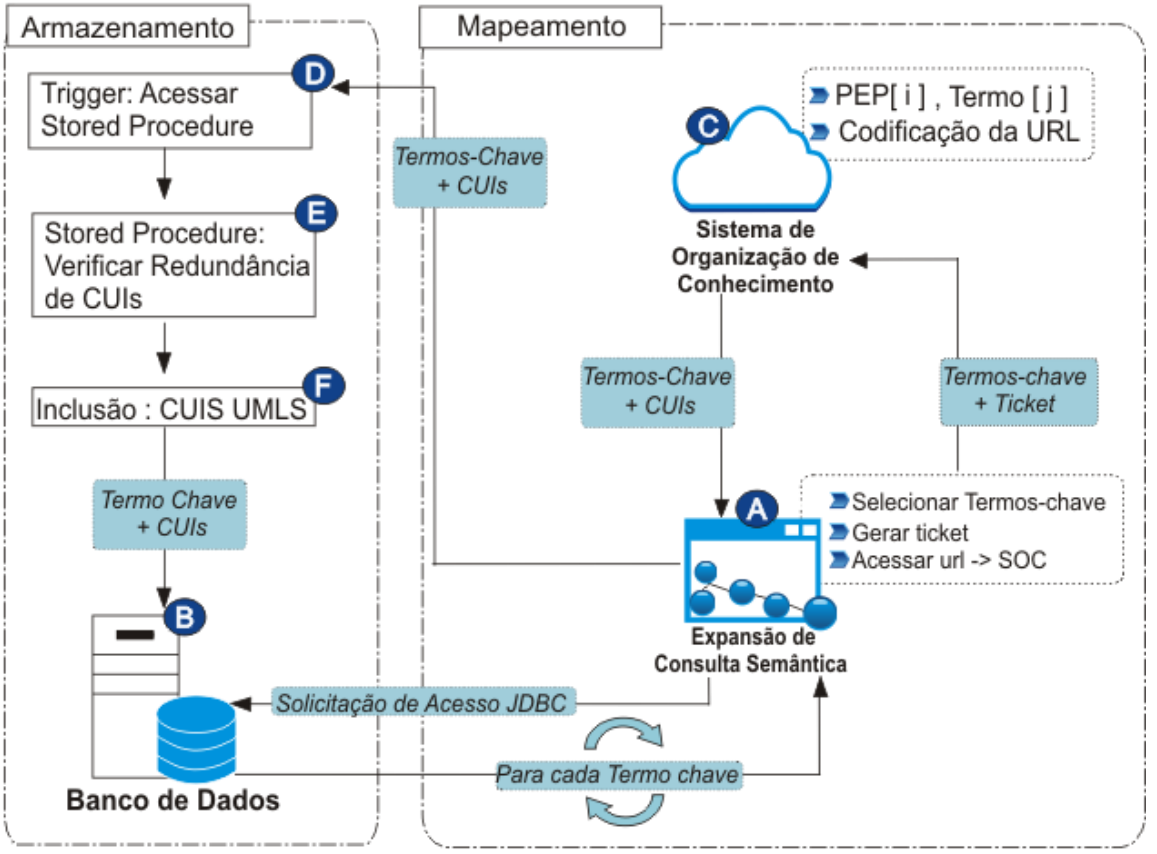

Figura 5.3: Componente de expansão de consultas

[30] API desenvolvida em JAVA
[31] https://uts.nlm.nih.gov/home.html
[32] Conexão via *JDBC (Java Database Connectivity)* ao banco de dados

A entrada ao servidor UTS (termo + *tickets*) é codificada e validada para que o conceito relacionado ao CUI seja identificado nos vocabulários da UMLS (item C da Figura 5.3). Para cada entrada de termo no servidor UMLS, vários CUIs podem ser retornados. A Tabela 5.1 apresenta um exemplo extraído da UMLS dos CUIs relacionados aos termos "Bactéria" e "Tosse".

Tabela 5.1: Exemplo de ocorrência de CUIs relacionados aos termos dos PEPs

Bactéria		**Tosse**	
CUI	Descrição	CUI	Descrição
D001422	Aderência Bacteriana	**10011229**	Tosse não produtiva
D058491	Carga Bacteriana	**10034738**	Tosse convulsa
D018410	Pneumonia Bacteriana	**D014917**	Coqueluche
D003773	Placa Dentária	**10011225**	Tosse diminuída
10027202	Meningite bacteriana	**10036790**	Tosse produtiva
10034133	Resistência microbiana	**10053779**	Tosse alérgica
10065198	Cistite bacteriana	**10003553**	Asma
10004016	Diarreia bacteriana	**11199725**	Tosse Crônica

Tecnicamente, cada ocorrência de termo-chave com os CUIs identificados na UMLS dispara uma *trigger* (item D da Figura 5.3) que realiza uma chamada para acessar um procedimento - *stored procedure* (Item E da Figura 5.3) - passando como parâmetro o código CUI relacionado com cada termo-chave. O procedimento verifica as duplicidades dos CUIs antes de inseri-los na base de dados. O objetivo é montar uma sub-base de termos semânticos que sejam relevantes para o conteúdo existente dos PEPs.

5.1.3. Indexação de PEPs

Neste componente (item F da Figura 5.1) utilizamos a plataforma de indexação *Lucene Apache Solr*. Ela dispõe de um módulo de configuração (item A da Figura 5.4) através de arquivos *XML* para apoiar os critérios usados no processo de indexação. Os PEPs armazenados na base de dados (item C da Figura 5.4) são importados[33] para o *Solr*. As bibliotecas e *plug-ins* apresentadas no (item A da Figura 5.4) permitem que o aplicativo por meio da *interface* gráfica (item B Figura 5.4) utilize os parâmetros pré-configurados e realize uma conexão ao banco de dados (item C da Figura 5.4) para dar início ao processo de indexação dos PEPs. Esse processo é realizado apenas uma vez devido ao fato do *Solr* ter seu próprio local de armazenamento de dados.

[33] "*MYSQL* conector 5.1" é usado para estabelecer a conexão.

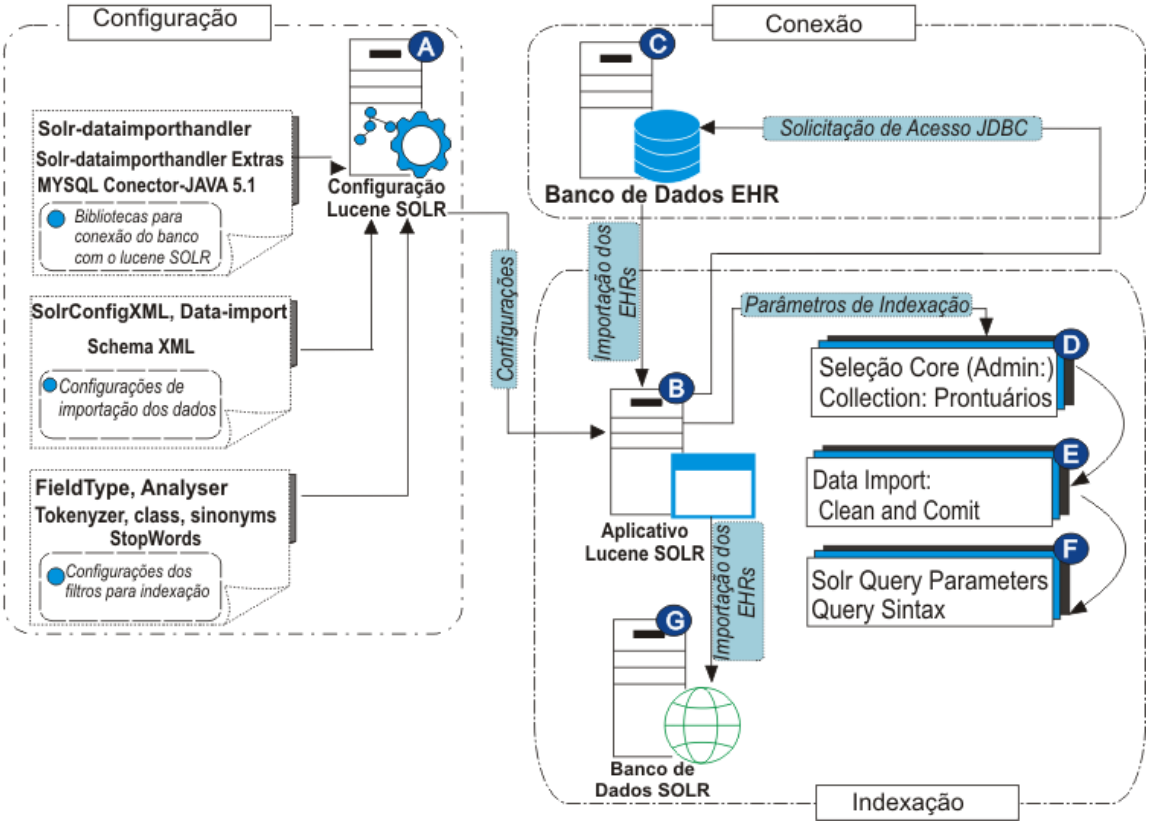

Figura 5.4: Componente de indexação de PEPs

Mais especificamente, para a indexação dos PEPs, utilizou-se o *collection* prontuários (item D da Figura 5.4) que foi previamente configurado para importar somente os campos de interesse dos PEPs. Neste ponto, é importante ressaltar que a configuração para indexação dos PEPs no *SOLR* seguiu configurações de filtros(Tokenizers)[34] padrões, exceto o filtro(*ASCIIFoldingFilterFactory*) para poder trabalhar com caracteres ASCII. Desta forma, a recuperação de documentos visou detectar documentos de acordo com a cadeia de caracteres e o número de ocorrências em que as palavras-chave ocorrem nos PEPs. Não foram utilizados filtros adicionais à configuração padrão, tais como MLT (*More Like This*) que é utilizado para detectar documentos semelhantes. Entendemos ainda que o *SOLR* é capaz de realizar buscas que vão além da simples comparação léxico-sintática.

O procedimento de importação de dados (item E da Figura 5.4) permitiu a indexação dos PEPs na base de dados do *Solr*. Adicionalmente, o módulo de pesquisas

[34] https://wiki.apache.org/solr/AnalyzersTokenizersTokenFilters

(item F da Figura 5.4) permite a entrada de consultas, *e.g.* (anamnese: “TOSSE”) AND (anamnese: “ALERGIA”) tendo como saída os PEPs que tiverem os termos “TOSSE” e “ALERGIA” indexados conforme a configuração do *Solr*.

5.1.4. Processamento de Consultas

O componente de processamento de consulta[35] (item G da Figura 5.1) refere-se ao núcleo do SiRBI, ou seja é onde está sua principal funcionalidade e diferencial. O módulo de processamento de consulta (item A da Figura 5.5) consiste da interface de usuário para a formulação de consultas. Ela permite ao usuário entrar com uma ou mais palavras-chave com suas respectivas dimensões e valores. Esses elementos refletem a necessidade do usuário no momento da consulta. A consulta é a entrada para o módulo de processamento de consulta pragmática (Item B da Figura 5.5). Para obter o resultado de busca com os PEPs ordenados, o módulo envolve um processo que considera um método de busca sintática, busca expandida e busca pragmática em sequência.

Busca Sintática (item C na Figura 5.5): As palavras-chave definidas na consulta são usadas na execução da ferramenta de busca sintática do *Solr* (item D da Figura 5.5). Por exemplo, se “Tosse” e “Alergia” são as respectivas palavras-chave, o índice é consultado verificando a ocorrência de PEPs que contenham esses termos de entrada da consulta. A ferramenta de busca sintática (*Solr*) retorna uma lista[36] de PEPs ordenados. A lista de PEPs resultante é armazenada na base de dados para ser utilizada nos próximos passos.

Busca Expandida (item E na Figura 5.5): A execução do método de busca expandida explora os termos relacionados com as palavras-chave definidas na consulta. Para cada palavra-chave, todos termos (identificados pelos CUIs) semanticamente relacionados são considerados (Item F da Figura 5.5). O módulo de busca sintática é chamado para cada combinação possível de palavra-chave, isto é, realiza-se a busca para todo o produto cartesiano dos conjuntos de termos relacionados. Os resultados da busca sintática são agregados e armazenados na base como um resultado intermediário.

[35] Desenvolvido na linguagem PHP
[36] Em formato JSON *(JavaScript Object Notation)*

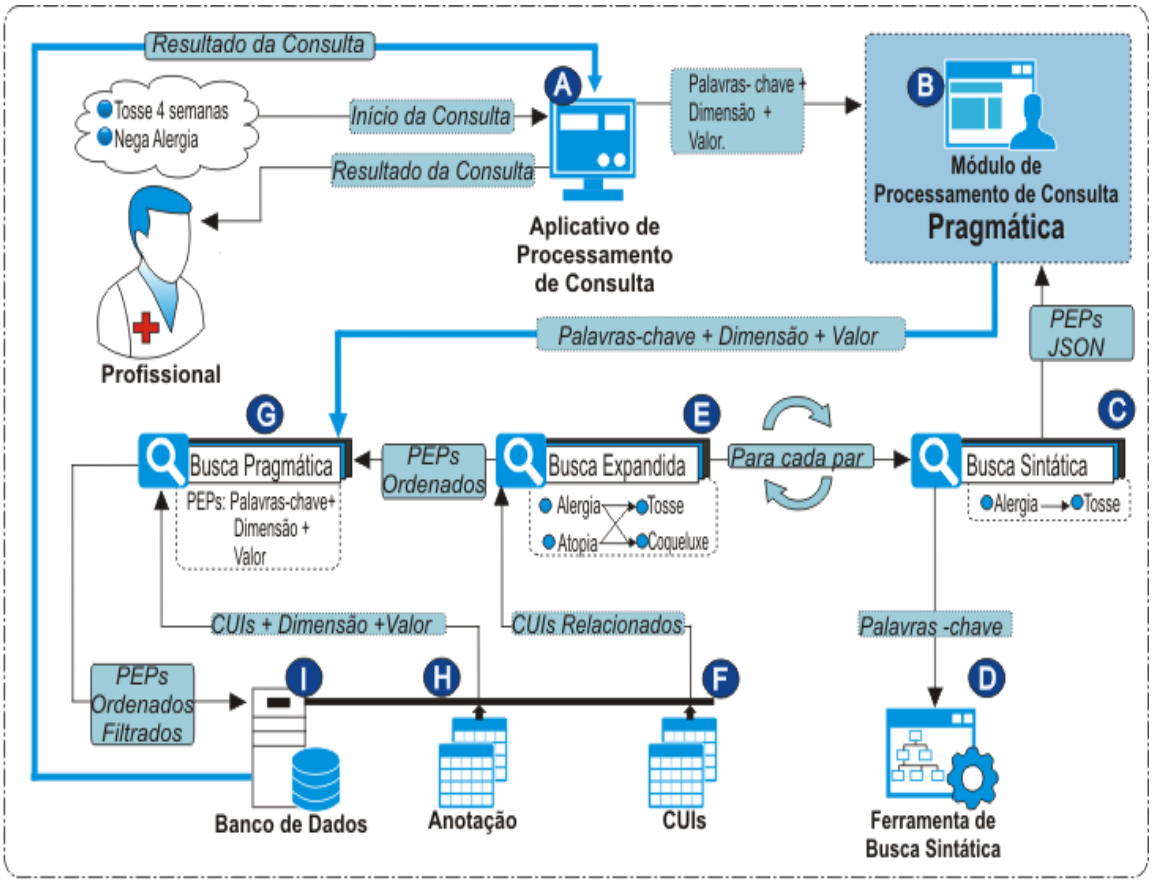

Figura 5.5: Componente de processamento de consulta pragmática

Busca Pragmática (Item G na Figura 5.5): Esta visa atuar como um filtro no resultado obtido pela busca expandida, aprimorando assim a precisão dos resultados de busca. Para este fim, a busca pragmática explora o conceito de anotação de ilocuções. Uma anotação consiste em efetuar uma marcação explícita entre um elemento textual, a dimensão envolvida naquele elemento e o valor associado a ela. A tabela 5.2 apresenta exemplos de anotação.

Tabela 5.2: Anotação Manual das Ilocuções em PEPs

1	***Dor lombar baixa --- (M545)***
	Paciente refere lombalgia, principalmente ao movimentar as pernas. Refere irradiação para o abdome. Nega febre. Nega sintomas urinários ou gastrointestinais. Nega uso de medicamentos. Nega uso diário de medicamentos. Nega atopias medicamentosas. o exame: Abdome semi-globoso, normotenso, sem massas ou vísceras palpáveis, doloroso a palpação profunda, difusamente. DB e

	Murphy negativos. Giordano negativo.			
2	**Classificação das Ilocuções**			
	Trecho Analisado	**Dimensão**	**Valor**	**Classe**
I_1	...*REFERE* ***lombalgia, principalmente ao movimentar as pernas. Refere irradiação...***	*Invenção (descritiva)*	Refere	**Asserção**
		Tempo (presente)	Refere	
		Modo (denotativo)	Refere	
I_2	...*NEGA* ***febre. Nega sintomas urinários ou gastrointestinais...***	*Invenção (descritiva)*	Nega	**Asserção**
		Tempo (presente)	Nega	
		Modo (denotativo)	Nega	
I_3	...*AFIRMA* ***o exame: Abdome semi-globoso, normotenso, sem massas ou vísceras palpáveis*** ...	*Invenção (prescritiva)*	Afirma	**Asserção**
		Tempo (presente)	Afirma	
		Modo (denotativo)	Afirma	

No estágio deste trabalho, as anotações foram efetuadas de maneira manual de modo a viabilizar as avaliações experimentais da pesquisa. O conjunto de anotações obtido para a base de PEPs foi realizada juntamente com profissionais da saúde.

A última etapa do componente de processamento de consultas consiste em confrontar os PEPs resultantes da busca expandida com o conjunto de anotações existente (item H da Figura 5.5). O algoritmo considera inicialmente os parâmetros da consulta declarada (item A da Figura 5.5) e percorre as anotações em busca de ocorrência nas anotações que estejam de acordo com os elementos da consulta. Por exemplo, se a consulta especifica uma busca com a palavra-chave "*tosse*" na dimensão "*tempo*" com o valor de "*4 dias*", o sistema procura nas anotações dos PEPs resultantes da busca expandida se existe alguma notação que associa o termo "*tosse*" com "*4 dias*" na dimensão "*tempo*".

Caso exista algum PEP no conjunto de resultados da busca expandida que contenha palavras-chave da consulta nas anotações, o sistema filtra o respectivo PEP. Tendo em vista que os PEPs já estão ordenados pelo componente de acordo com a

ferramenta de busca sintática (item D da Figura 5.5), à medida que o algoritmo encontra os PEPs, ele separa os PEPs filtrados (item G da Figura 5.5). Logo, o algoritmo retorna para o usuário os PEPs ordenados e filtrados de acordo com os elementos especificados da consulta de entrada.

5.2. Interfaces de Usuário e Funcionamento do Sistema

Nesta seção são apresentadas as principais interfaces e o funcionamento do sistema, que implementa a arquitetura e funcionalidades descritas na seção anterior. A Figura 5.6 apresenta a interface de formulação de consulta. Nela o usuário pode declarar os parâmetros de consulta que refletem sua necessidade de informação. Através dela, o usuário insere parâmetros de consulta simples ou compostos por mais de uma combinação de palavras-chave, dimensões e valores, conforme destacado a seguir:

- **Palavra-chave da consulta:** campo que o usuário entra com termos associados à patologia, doença ou qualquer outro termo que julgue importante para a seleção de PEPs. Para este fim, o sistema implementa uma técnica[37] que a cada letra digitada pelo usuário, é realizada uma consulta aos termos base e apresenta sugestões em uma caixa de seleção dinâmica;
- **Dimensão:** este campo consiste numa caixa de seleção de valores onde o usuário pode optar por escolher uma das 3 dimensões ("*Tempo*", "*Modo*" e "*Invenção*"). Ressalta-se que a cada adição de uma nova linha de consulta, uma nova "Dimensão" pode ser inserida para compor uma consulta;
- **Valor:** Este campo permite ao usuário inserir um valor que caracteriza a dimensão da ilocução relacionada com a palavra-chave (*e.g., "4 dias"* caracteriza a dimensão "*tempo"* para a palavra-chave "*tosse"*).

O exemplo ilustrado na Figura 5.6 apresenta uma consulta composta por quatro partes, onde cada campo sucessor tem correlação com seu campo predecessor:

1-	**Tm**	< Tosse não Produtiva>	+	**Dm**	<Tempo>	+	**Vl**	<4 Dias>
2-	**Tm**	<Alergia>	+	**Dm**	<Invenção>	+	**Vl**	<Refere>
3-	**Tm**	<Febre>	+	**Dm**	<Invenção>	+	**Vl**	<Refere>
4-	**Tm**	<Penicilina>	+	**Dm**	<Tempo>	+	**Vl**	<2 Dias>

[37] Explora a tecnologia AJAX (*Asynchronous Javascript and XML*)

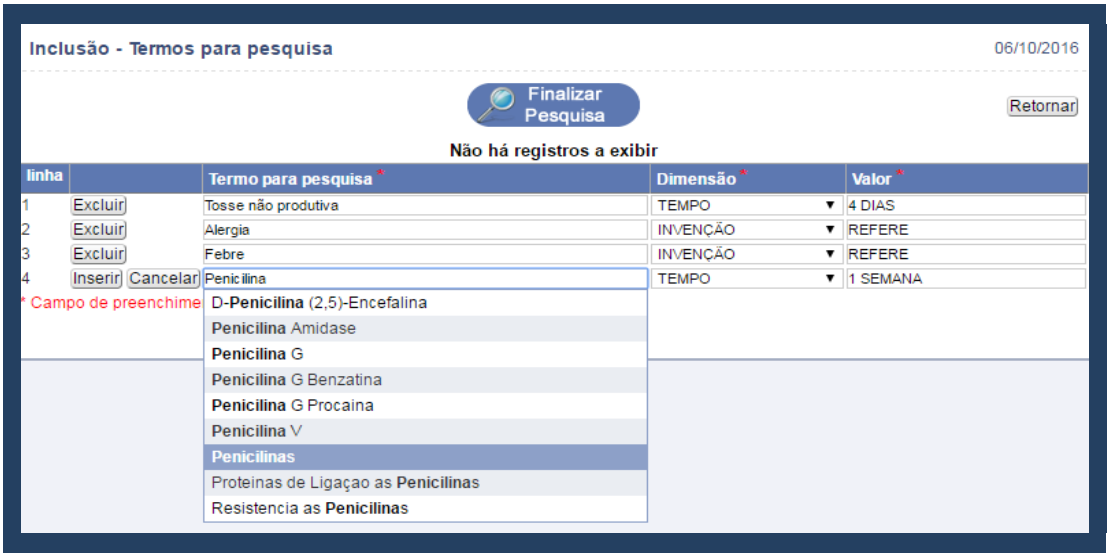

Figura 5.6: Interface de Formulação de Consultas

A pesquisa é executada através do elemento de interface "Finalizar Pesquisa". Ao término do processamento da consulta, uma interface de resultados da busca é apresentada ao usuário (Figura 5.7). Os resultados são exibidos em uma lista ordenada separados por uma quebra de linha (destacado em verde na Figura 5.7). A primeira quebra de linha (Recuperação Pragmática) inicialmente apresenta PEPs que contenham os termos, dimensões e valores exatamente iguais aos da entrada da consulta. Já a segunda quebra de linha (Recuperação Pragmática Aproximada) retorna ao usuário PEPs que contenham conceitos do domínio relacionados a consulta definida.

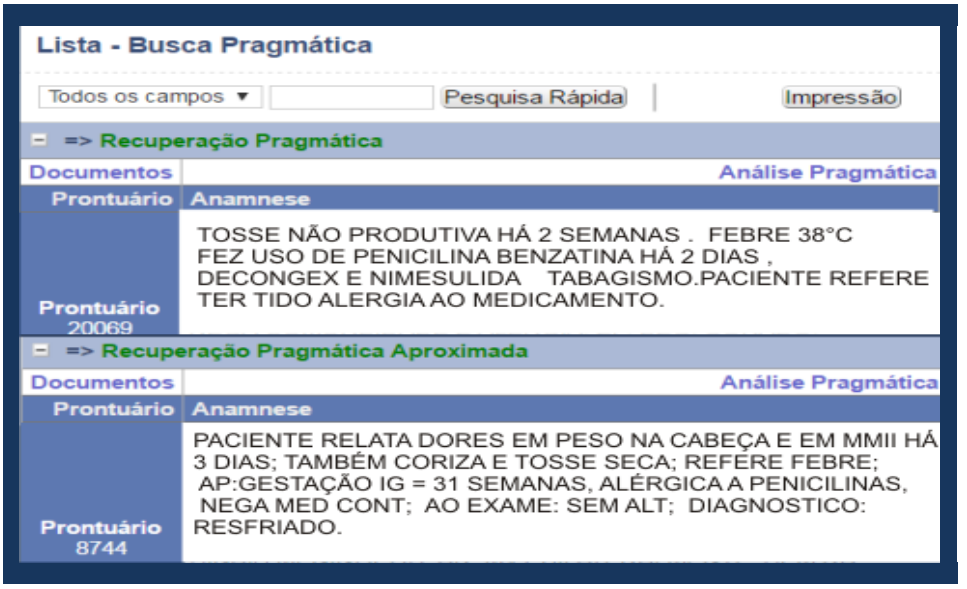

Figura 5.7: Interface de Resultado da Consulta (exemplo de resultado)

5.3. Síntese do Capítulo

Este capítulo apresentou a implementação do método de recuperação de informação baseado em intenções em um protótipo funcional. Descrevemos o SiRBI (Sistema de Recuperação com Base em Intenções) e detalhamos os componentes envolvidos no desenvolvimento do sistema em sua arquitetura.

A implementação foi repleta de desafios, as inovações providas pelo sistema requereram o uso de uma ampla gama de tecnologias e ferramentas. Primeiramente, foi preciso lidar com os desafios de processamento e expansão de consultas na língua portuguesa. Implementamos componentes específicos para processar o conteúdo dos PEPs e adotamos o uso do UTS para apoiar a expansão de consultas. A implementação da indexação de PEPs via o *Solr* foi desafiadora, visto que foi necessário lidar com diversos aspectos de configuração e realizar testes extensos para permitir a indexação adequada dos PEPs e averiguar seu funcionamento correto.

As anotações de ilocuções desenvolvem um papel chave no comportamento e resultados do algoritmo de recuperação. Um mecanismo de anotação inteiramente automático ainda é um desafio em aberto de pesquisa. Contudo, nesta pesquisa constate-se que o processo de anotação manual demonstrou-se factível aos cenários abordados, permitindo que o sistema pudesse se beneficiar das anotações e da adoção de intenção do usuário como parte da recuperação de PEPs. Finalmente, as principais funcionalidades do sistema foram descritas incluindo as interfaces de declaração de consulta e resultado de busca.

Capítulo 6

Avaliação Experimental

Este capítulo apresenta a avaliação experimental desta pesquisa com o objetivo de testar a proposta em um contexto real de recuperação de informação. Para este fim, exploramos cenários de busca apontados por médicos e usamos o sistema SiRBI para recuperar os PEPs nesses cenários. A Seção 6.1 apresenta o design do experimento incluindo os participantes, o procedimento conduzido, métricas, bem como os cenários e a importância deles do ponto de vista médico. Enquanto a Seção 6.2 detalha os resultados obtidos, a Seção 6.3 discute os benefícios e limitações da investigação. Finalmente, a Seção 6.4 sintetiza as contribuições do capítulo.

6.1. Design do Experimento

O estudo experimental foi realizado no contexto da área de saúde e os dados explorados foram coletados de um PA situado na cidade de "Águas de Lindóia", SP, Brasil. O objetivo da avaliação foi medir os benefícios da busca pragmática na recuperação de informação em PEPs. As seguintes etapas estão envolvidas no experimento:

1. Apresentação dos Participantes envolvidos
2. Definição dos cenários no contexto médico
3. Anotação manual de ilocuções
4. Elaboração da base de referência para testes
5. Configurações do experimento
6. Definição das métricas exploradas

Participantes. O experimento foi realizado em conjunto com dois profissionais da área de saúde com especialidades e formação distintas. O primeiro profissional é um médico, especialista em alergologia há 22 anos, e o segundo é uma enfermeira, especialista em ginecologia e obstetrícia com 18 anos de experiência. Esses profissionais, envolvidos diretamente nas tarefas, foram auxiliados por outros profissionais que atuaram de maneira secundária no experimento, sendo: 1 enfermeiro

com 6 anos de experiência, 1 enfermeira com 5 anos de experiência e 1 técnica de enfermagem com 26 anos de experiência. Neste contexto, ressalta-se que os participantes desta avaliação não são os mesmos profissionais que realizaram as anotações das ilocuções nos documentos médicos, para se evitar um viés sobre os participantes. Os cinco participantes determinaram os cenários de busca relevantes no domínio e classificaram os PEPs para formar uma base de referência com a qual os resultados obtidos pelo sistema foram contrapostos.

Definição dos cenários. Para a definição dos cenários, primeiramente foi realizada uma análise geral exploratória no conjunto de PEPs disponíveis na base (13.300 registros). A Figura 6.1 apresenta os participantes interagindo na definição dos cenários.

Figura 6.1: Análise de PEPs para definição de cenários

A análise apontou que a hipótese diagnóstica (um código CID10) associada a cada PEP poderia ser um meio de selecionar e restringir um subconjunto relevante de PEPs para o estudo. Isto se deve ao fato que a hipótese diagnóstica está ligada ao contexto da anamnese do paciente. A (*cf.,* Tabela 4.1) apresenta as hipóteses diagnósticas com maior índice de ocorrências decorrente da análise realizada nos PEPs da base selecionada.

Através das hipóteses diagnósticas, os profissionais puderam elaborar perguntas sobre quais gostariam de obter respostas mais precisas segundo informações contidas na base. Dois cenários demostraram grande potencialidade na realização desta avaliação experimental, sendo eles:

Cenário 1: Neste cenário o médico gostaria de recuperar registros de pacientes que relatem tosse no período noturno mas que não apresentem qualquer tipo de alergia.

A detecção da intensidade da tosse noturna pode ajudar a identificar o agente causador (poeira, fungos, ácaros, umidades, restos de insetos, entre outros) que se instalam e se proliferam no ambiente residencial. Nesse caso, a conduta médica entra com iniciativas para combater os agentes causadores. Na definição da consulta no sistema, este cenário é composto por duas sentenças conjuntivas, sendo elas:

1- **Tm** < Alergia> + **Dm** <Modo> + **Vl** <Nega>

2- **Tm** <Tosse> + **Dm** <Tempo> + **Vl** <À noite>

Cenário 2: Neste cenário o médico investiga PEPs nos quais os pacientes relatam a ausência de febre, refere não ser alérgico e somente ter tosse. A combinação de elementos deste cenário é relevante, pois a febre pode ser um sinal de alerta de uma doença que precisa ser investigada com maior rapidez. Ela abre a investigação para outras patologias de origem viral, bacteriana e fúngica. O cenário abordado a tosse não está associada com alergia, o que leva o profissional à investigação de outros parâmetros que identifiquem a causa da tosse. Além disso, pela ausência de febre é possível evitar submeter o paciente a tratamentos mais severos como antibióticos, que ao longo do tempo causam resistências no organismo. Este cenário é composto por três sentenças conforme o seguinte:

1- **Tm** < Tosse> + **Dm** <Invenção> + **Vl** <Refere>

2- **Tm** <Alergia> + **Dm** <Modo> + **Vl** <Nega>

3- **Tm** <Febre> + **Dm** <Modo> + **Vl** <Nega>

Anotação das ilocuções. O mecanismo de busca pragmática se fundamenta em anotações sobre instâncias de ilocuções relacionadas ao conteúdo dos PEPs. Para viabilizar esta avaliação experimental, foi desenvolvida uma funcionalidade no sistema para apoiar o usuário a inserir as anotações sobre ilocuções a partir do conteúdo dos PEPs.

Para não exigir que toda a base de registros fosse anotada, consideramos o subconjunto com base na classificação de hipóteses diagnósticas que se enquadram os cenários estudados. A Figura 6.2 ilustra o formulário no sistema *SiRBI* para anotação e

classificação manual dos PEPs, ela é necessária para o funcionamento da busca pragmática.

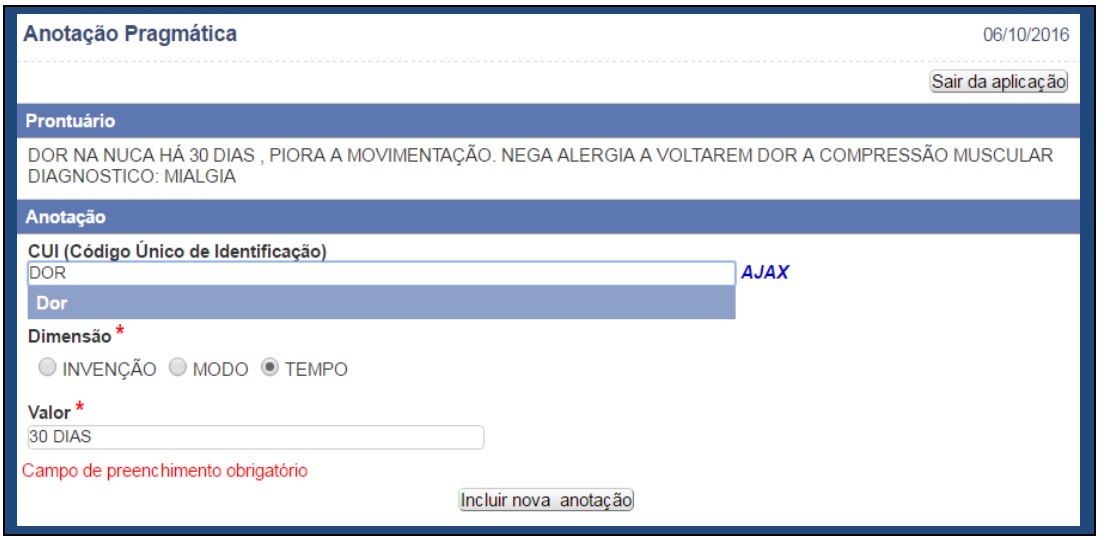

Figura 6.2: Interface para anotação pragmática

A interface (topo da Figura 6.2) apresenta a descrição de um PEP em linguagem natural para análise das ilocuções que está organizada em três campos distintos, sendo eles:

1) **CUI (Codigo Único de Identificação):** Tem a finalidade de associar um elemento textual com CUIs existentes na UMLS. Para cada trecho analisado ou termo-chave, o usuário pode classificar um CUI, *e.g*, "Dor", "Nauseas", "Febre", *etc*.

2) **Dimensão:** A anotação segue de acordo com o cubo de classificação de ilocuções adotado nesta pesquisa. O usuário pode optar por três dimensões "*Tempo*", "*Modo*" ou "*Invenção*". A Figura 6.2 exibe um PEP que em determinado trecho do texto apresenta um CUI relacionado a "*dor na nuca*" seguido de uma dimensão de tempo "*há 30 dias*". Neste caso a seleção da dimensão está relacionada ao *"Tempo"*.

3) **Valor:** O último campo de anotação tem como finalidade instanciar as classes de ilocuções. Nele o usuário pode introduzir um valor relacionado à dimensão de acordo com a interpretação do PEP. No exemplo ilustrado na Figura 6.2, tem-se o campo preenchido como "*30 dias*" devido ao fato do campo "*dimensão*" estar relacionado ao "*Tempo*".

Elaboração da base de referência. Esta etapa visou classificar a relevância dos PEPs em um subconjunto da base experimental para cada um dos cenários em estudo. A

construção de um conjunto de referência é necessária por não haver um conjunto padronizado de teste no contexto desta pesquisa. Os participantes foram convidados a analisar o conteúdo dos PEPs de uma parte da base para indicar o grau de importância de cada PEP para o objetivo (de recuperação de informação) do cenário. Para a elaboração do subconjunto de PEPs foi considerado o resultado da busca expandida, que envolve a busca dos PEPs com todas as variações possíveis de palavras-chaves relacionadas no UMLS. Apenas os dois participantes principais atuaram na elaboração da base de referência, um para cada cenário.

A Figura 6.3 apresenta a funcionalidade elaborada no sistema para apoiar a análise de relevância dos prontuários pelos participantes. Através dela, o usuário qualifica os PEPs de acordo com os cenários como irrelevante, pouco relevante, relevante e muito relevante. Para tanto, conforme apresenta Figura 6.4, ele deve clicar no *checkbox* correspondente na linha do PEP analisado.

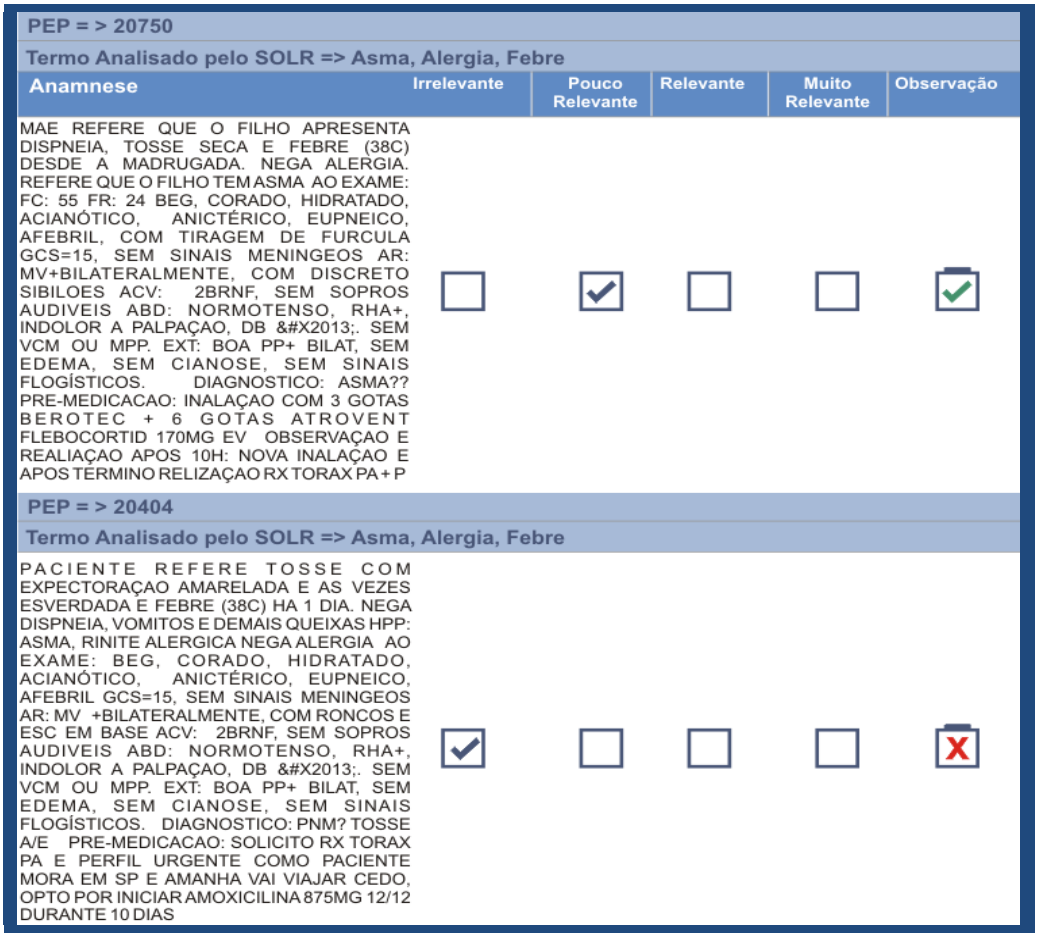

Figura 6.3: Avaliação da Relevância dos PEPs

Como exemplo, a Figura 6.3 apresenta o PEP com ID '20750' avaliado pelo profissional de saúde como um PEP "Pouco Relevante" (para o cenário de busca em estudo) e o PEP com ID '20404' como "Irrelevante". A interface apresentada na Figura 6.3 possibilita ainda a inclusão de uma observação sobre a análise de cada PEP. Nela o usuário pode inserir informações que justifiquem a avaliação ou outros dados que julguem importantes no contexto do PEP avaliado.

Configurações do experimento. Para verificar os benefícios do mecanismo de busca pragmática desenvolvido no SiRBI, cada cenário foi testado explorando quatro configurações: (1) – Busca *Solr*; (2) - Busca "Sintática" SiRBI[38]; (3) Busca Expandida e (4) Busca Pragmática. Utilizando as bases de referência definidas para os cenários, o objetivo foi verificar a capacidade da busca pragmática do SiRBI melhorar, por meio das anotações e algoritmo proposto, os resultados obtidos na busca sintática e expandida. Não é objetivo promover uma competição exaustiva entre as abordagens existentes. Ou seja, estamos em um primeiro momento interessados em analisar o "delta" obtido entre as configurações (2) e (3) em relação à configuração (4), verificando a existência de benefício da abordagem (isolando o efeito) quanto utilizamos o algoritmo em conjunto com ferramentas existentes, tais como o *Solr*. Já a configuração (1), de forma complementar, foi introduzida para ilustrar em termos comparativos o comportamento do mecanismo de Busca *Solr* quando diretamente implementado e utilizado na interface de administração (Busca *Solr*), pois esta possui filtros e extensões adicionais aos utilizados na configuração (2). Para este fim, as consultas foram efetuadas usando as palavras-chaves em adição daos termos utilizados na configuração (2), essas palavras chaves incluem os termos relacionados ao valor associado a cada dimensão; possibilitando assim obter uma melhor precisão de busca[39]. Por exemplo, uma consulta inclui na Busca *Solr*:("Tosse" *AND* "A NOITE" *AND* "ALERGIA" *AND* "NEGA"), enquanto na Busca Sintática-SirBI explora apenas as palavras-chaves definidas pelo usuário no envio de consultas para o *Solr* ("Tosse" AND "Alergia"). A Busca Expandida (configuração 3) considera a expansão semântica de termos pelo *UMLS* sobre os resultados da configuração (2), conforme definido no algoritmo PraSA.

[38] Embora o *Solr* seja capaz de realizar buscas que vão além da simples comparação léxico-sintática, o termo "sintática" foi utilizado para diferenciar da busca expandida e pragmática proposta neste trabalho.

[39] Estes parâmetros adicionais não foram utilizados na configuração (2), pois temos o objetivo no algoritmo de obter a maior cobertura possível com as buscas sintáticas e expandidas.

Métricas de avaliação. Para possibilitar uma comparação objetiva entre as configurações experimentais, exploramos medidas padrões da área de recuperação de informação utilizadas para avaliar a eficácia de mecanismos de busca. As medidas exploradas foram as de Precisão, Cobertura e Medida-f.

Primeiramente, a Precisão é dada pela soma daqueles PEPs recuperados pelo mecanismo de busca e considerados 'Relevante' ou 'Muito Relevante' pelo especialista dividido pelo Total de PEPs recuperados pelo mecanismo em análise (resultado da execução de cada configuração), conforme a seguinte formula:

$$Precisão = \frac{(PEPs\ Relevantes + PEPs\ Muito\ Relevantes)}{(Total\ de\ PEPs\ recuperados\ pelo\ mecanismo\ em\ análise)}$$

A medida de Cobertura avalia a fração dos documentos que são definidos como 'Relevantes' e 'Muito relevantes' para a consulta que é recuperada com sucesso. Assumiu-se os resultados da busca expandida como o conjunto total possível de PEPs recuperados (conjunto que foi avaliado pelos participantes), pois ela contém a expansão de todos os CUIs. Logo, consideramos como divisor o número de PEPs classificados como 'Relevante' ou 'Muito Relevante' nos resultados da Busca Expandida. Assim, a cobertura é medida na execução de cada configuração experimental pela seguinte formula:

$$Cobertura = \frac{(PEPs\ Relevantes + PEPs\ Muito\ Relevantes)}{(PEPs\ Relevantes + PEPs\ Muito\ Relevantes\ da\ Busca\ expandida)}$$

A Medida-f define a média harmônica entre Precisão e Cobertura.

$$Medida\text{-}F = 2 * \frac{(Precisão * Cobertura)}{(Precisão + Cobertura)}$$

6.2. Resultados

Nesta seção, apresentamos os resultados obtidos do experimento conduzido com os cenários e as configurações definidas para os cenários a seguir.

Cenário 1. A Tabela 6.1 apresenta os resultados obtidos para o cenário 1 considerando Busca Sintática-SiRBI (palavras-chave), Busca *Solr* (palavras-chave, dimensões e valores), Busca Expandida (usando o *UMLS*) e a Busca Pragmática (proposta nesta investigação). Esta análise avaliou um subconjunto de 308 PEPs de um

total de 13.300. A Busca Sintática-SiRBI retornou um total de 239 PEPs, dos quais 10 são considerados 'Relevantes' e 25 'Muitos Relevantes', o que resulta em *0,15* de Precisão (15%). Assim, a Busca Sintática retornou 35 resultados (entre 'Relevantes' e 'Muitos Relevantes') de um total de 37 resultados ('Relevantes' e 'Muitos Relevantes' da Busca Expandida na Tabela 6.1), resultando em *0,95* de Cobertura (95%). Portanto, a Busca Sintática-SIRBI obteve *0,25* de Medida-F.

Ao considerar termos, dimensões e valores a Busca *Solr* retornou um total de 24 PEPs, aproximadamente 10% do total de ocorrências da Busca Sintática-SiRBI. Desta forma a cobertura foi relativamente mais baixa 43%, porém a precisão dos resultados obteve um ganho de 52% ao considerar as dimensões e valores no momento da recuperação de PEPs se comparado com a Busca Sintatica-SiRBI que somente analisou somente as palavras-chave do cenário. A Busca *SOLR* resultou em uma Medida-F de 52%.

Tabela 6.1: Resultados para o Cenário 1

	Irrelevante	Pouco Relevante	Relevante	Muito Relevante	TOTAL	Precisão	Cobertura	Medida-F
Busca *Solar*	3	5	5	11	24	0,67	0,43	0,52
Busca Sintática - SiRBI	25	179	10	25	239	0,15	0,95	0,25
Busca Expandida	72	199	12	25	308	0,12	1,00	0,21
Busca Pragmática	0	2	8	25	35	0,94	0,89	0,92

Enquanto a Busca Expandida tem *1,00* de Cobertura (100%), a Precisão é levemente mais baixa que a Busca Sintática-SiRBI (diferença de 3%), quando comparada com a Busca *Solr* os resultados se acentuam com uma diferença de 55%.

Já a Busca Pragmática retornou um total de 33 PEPs, com *0,89* de Cobertura (89%). Embora a Cobertura seja menor que a Busca Sintática-SiRBI e a Busca Expandida, nota-se uma melhoria significativa ao considerar a relação entre a intenção e os conceitos nas anotações em PEPs. Logo, há um ganho evidente na medida de Precisão, que subiu para *0,94* (94%). Resultado em uma melhor Medida-F com 0,92. Isso sugere uma melhoria na qualidade dos resultados proporcionados pelos filtros da Busca Pragmática ao serem aplicados nas buscas executadas na configuração (2) e (3) do

experimento para este cenário. Vale ressaltar que a busca pragmática também obteve Medida-F superior à configuração (1).

Cenário 2. Este cenário explora uma consulta mais complexa do que no cenário 1 envolvendo a combinação de mais elementos e conceitos. A Tabela 6.2 apresenta os resultados obtidos em uma análise de um subconjunto de 263 PEPs. A Busca Sintática-SiRBI apresenta uma Precisão de *0,47* (47%), ou seja, ela retornou 96 resultados entre 'Relevantes' e 'Muito Relevantes' de um total de 203 PEPs recuperados. Esta configuração obteve *0,89* de Cobertura (89%), pois foram retornados 96 dos 108 resultados 'Relevantes' e 'Muito Relevantes' da Busca Expandida. Ao considerar as palavras-chave, dimensões e valores a Busca *Solr* obteve melhores resultados para a precisão 57% e cobertura 91% em relação a Busca Sintática-SiRBI. Entretanto, o resultado para a Medida –F foi de apenas 3% maior, obtendo 65%.

Assim como foi observado no primeiro cenário, a Busca Expandida apresenta uma leve queda na Precisão comparada com a Busca Sintática (de 0,47 para 0,41). Embora a Busca Pragmática tenha obtido resultados inferiores aos alcançados no cenário 1, há uma melhoria substancial na Precisão dos PEPs ao utilizarmos os filtros da Busca Pragmática, sendo *0,78* obtidos na Busca Pragmática, *0,41* da Busca Expandida e *0,47* da Busca Sintática. Apesar da Cobertura da Busca Pragmática ser inferior (*0,73*) comparada todas outras configurações de busca, a Medida-F obtida (*0,76*) supera as demais configurações.

Tabela 6.2: Resultados para o Cenário 2

	Irrelevante	Pouco Relevante	Relevante	Muito Relevante	TOTAL	Precisão	Cobertura	Medida-F
Busca *Solar*	45	51	44	54	194	0,51	0,91	0,65
Busca Sintática - SiRBI	53	54	42	54	203	0,47	0,89	0,62
Busca Expandida	90	65	49	59	263	0,41	1,00	0,58
Busca Pragmática	3	19	34	45	101	0,78	0,73	0,76

6.3. Discussão

Os resultados experimentais obtidos indicam o potencial do SiRBI ao considerar as intenções dos usuários na recuperação de informação em repositórios de PEPs. Os resultados apresentam uma melhora significativa da Busca Pragmática em comparação com a Busca Expandida e a Busca Sintática.

No cenário 1, a consulta testada foi específica ao considerar pacientes que referem "tosse à noite" e "negam alergia". A Busca Sintática recuperou PEPs que continham apenas os termos "tosse" e "alergia" independentemente dos parâmetros relativos à dimensão e valor das ilocuções. Em contrapartida, a Busca *Solr* considerou os valores das dimensões para cada cenário.

Os resultados dos PEPs avaliados como "Irrelevantes" e "Pouco Relevantes" se devem ao fato da Busca Sintática-SiRBI e Busca SOLR ter recuperado PEPs que continham os termos "tosse e alergia", porém que divergiram em relação à dimensão e seu valor na consulta. Por exemplo, PEPs que "referem" tosse e que "possuem" algum tipo de alergia são retornados. Os resultados dos PEPs classificados como "Relevantes" e "Muito Relevantes" representam documentos que apresentaram os termos, dimensões e valores conforme a solicitação da consulta.

Os resultados recuperados pela Busca Expandida apresentaram alto índice de PEPs "Irrelevantes" e "Pouco relevantes" com uma Precisão mais baixa devido ao fato de mais combinações de termos do domínio serem consultados (*i.e,* ela considera a combinação de todos os termos relacionados a "tosse" e "alergia"). Isso apresenta uma tendência de piorar a Precisão e melhorar a Cobertura. A Busca Pragmática apresentou bons resultados para cenário o 1. Isso se deve ao fato de que esse cenário foi específico ao filtrar "pacientes que referem tosse à noite" e que "negam alergia". Desta forma, a Busca Pragmática permite eliminar PEPs que não expressam exatamente os parâmetros da consulta. Consequentemente, há uma melhoria no resultado de Precisão e Medida-F.

O cenário 2 propõe uma consulta mais complexa embora os termos "tosse", "alergia" e "febre" sejam mais fáceis de serem encontrados nos PEPs quando comparado a um termo mais específico como "tosse noturna". As avaliações apresentam alto índice de PEPs "Irrelevantes" e "Pouco Relevantes" tanto para a Busca Sintática-SiRBI e

SOLR, quanto na Expandida. A Busca Sintática-SiIRBI recuperou PEPs de acordo com os termos expressos na consulta.

Já a Busca *Solr* considerou cada dimensão e o valor relativo ao termo da consulta aumentando a cobertura em 2%, embora tenha usados mais palavras chaves conectadas pelo operador "AND". Isto ocorreu devido as extensões adicionais utilizadas na interface de busca do *Solr*, que não foram adotadas na Busca Sintática-SiRBI. O uso de tais extensões também tem o potencial de melhorar os resultados das configurações (2), (3) e (4). Ao utilizarmos extensões (como a MLT) com um número menor de parâmetros presentes na configuração (2), podemos aumentar a cobertura das configurações (2) e (3) e potencialmente da configuração (4). Entretanto, isso demandaria pesquisas adicionais sobre o entendimento e ajustes de extensões e configurações que está fora do escopo deste trabalho.

A Busca Expandida procurou recuperar PEPs relacionados aos termos expressos na consulta. Isso derivou no aumento de 60 resultados dos quais apenas 12 eram 'Relevantes', revelando uma baixa precisão do método de fazer o produto cartesiano entre os termos relacionados aos expressos na consulta.

Assim como no cenário 1, a Busca Pragmática demonstrou uma melhora de qualidade ao considerar a relação entre as intenções e conceitos nas anotações de PEPs, pois permite efetuar um filtro mais consistente nos resultados de acordo com os meta-dados disponíveis. Ela eliminou do resultado da Busca Expandida os PEPs que não estão relacionados com as dimensões e valores das ilocuções, permitindo uma seleção mais adequada dos resultados com uma perda "controlada" da Cobertura.

Os resultados obtidos nos levam a algumas reflexões sobre desafios a serem enfrentados em investigações futuras, tal como a busca por um método (semi-) automático para anotação das ilocuções. Embora o método elaborado tenha sido considerado factível do ponto de vista computacional e operacional (para "pequenas" bases de dados), o estudo de um anotador automático é desejável para o uso do método em larga escala, isto é escalar de dezenas de milhares de PEPs para milhões. Ao analisar a anotação manual é possível identificar certos padrões que repetem frequentemente, como o uso de verbos, expressões de tempo e relatos de pacientes. Isso pode ser um fator importante na construção de um mecanismo de anotação explorando ontologias do

domínio para a detecção das anotações e para aprimorar a qualidade da busca. Por exemplo, uma anotação considerando o valor “tosse à noite” é equivalente à “tosse noturna”. Um refinamento das anotações nesse sentido pode aprimorar a qualidade dos resultados da Busca Pragmática.

O uso de *tags* por quem produz o conteúdo também pode ajudar no desenvolvimento de aplicações em larga escala e na qualidade das anotações produzidas. Uma vez que o texto passa por uma interpretação de um terceiro no processo de anotação, seja ele um humano no caso da anotação manual ou um agente computacional no caso da anotação automática. Essa diferença de interpretação pode explicar parte da diminuição da cobertura na Busca Pragmática, *i.e.*, foram “filtrados” PEPs que o usuário julgam ser importantes.

Com relação à execução do algoritmo, embora este trabalho não esteja focado em *performance* no que diz respeito às medidas de tempo, notamos que o método aplicado deve ser otimizado. Para este estudo, os cenários foram executados em um computador com processador i5-3470 (3,20GHZ) dispondo de 2GB de memória RAM. Nesta configuração apresentou média de 25 minutos para executar o processamento da Busca Pragmática que inclui a execução das buscas Busca Sintática e Busca Expandida. Para o cenário 2, que possui mais parâmetros de consulta, foram necessários 60 minutos.

A maior demanda de tempo está relacionada ao processo de execução da Busca Expandida que explora as relações entre todos os termos relacionados através de um produto cartesiano. Assim, uma investigação entre as relações entre os termos expandidos da consulta poderia otimizar a performance de tempo, *e.g.*, para uma consulta entre os termos “tosse” e “alergia”, as relações entre eles poderiam ser exploradas no sentido de que nem toda alergia ocasiona tosse e nem toda tosse é ocasionada por algum tipo de alergia. Além disso, uma investigação mais detalhada do processo de expansão de consultas também pode influenciar positivamente nos resultados da medida de Precisão da Busca Expandida e da Busca Pragmática.

Trabalhos adicionais também devem ser conduzidos para investigar mais precisamente a ordenação dos resultados. O estágio atual implementado no sistema “herda” o modelo de ordenação da plataforma *Solr*. Por esse motivo não foram realizadas comparações adicionais entre as configurações do experimento. Um estudo

mais detalhado pode sugerir refinamentos na ordenação para lidar com aspectos que explore parâmetros da Busca Pragmática.

Um fator importante no que diz respeito às lições aprendidas na experimentação com usuários, se refere às dificuldades com o método de avaliação com os profissionais. Observamos que inicialmente os profissionais avaliaram o conteúdo dos PEPs de acordo com a interpretação geral e grau de impacto no quadro clínico do histórico do paciente (independente do cenário), e não se eles deveriam ser retornados pela busca conforme solicitado. Desta forma, por exemplo, se a Busca Expandida apresentasse PEPs fora do contexto dos cenários, *i.e*, os PEPs apresentassem conteúdos de interesses diversos (válidos em outros contextos), esses eram avaliados como 'Relevantes' e 'Muito relevantes'.

Embora isso levasse a medidas melhores de Precisão em todas as configurações, tal inconsistência foi detectada e esclarecido o papel da busca para os participantes. Isso nos leva a entender a necessidade de uma explicação inicial mais detalhada aos participantes em experimentos futuros. Um design de interface mais cuidadoso também pode evitar falsas expectativas dos usuários quanto ao papel do mecanismo busca. Faz-se necessária ainda experimentar outros cenários e uso real em longo prazo, para assim verificar de maneira completa os benefícios da Busca Pragmática, bem como elucidar alternativas para aprimorar a proposta.

Embora o protótipo SiRBI tenha demonstrado a necessidade de pesquisas adicionais em termos de anotações (semi-)automáticas, *performance* e *design* de *interfaces* para a expressão das consultas, os resultados dos experimentos empíricos são certamente avanços e demonstram o potencial do método ao considerar o uso explícito de intenções na recuperação de informação em registros médicos.

6.4. Síntese do Capítulo

Este capítulo apresentou a avaliação experimental da proposta do método de recuperação de informação implementado no sistema SiRBI. Por meio do experimento conduzido foi possível averiguar os benefícios da proposta na melhoria das medidas padrões para avaliar mecanismos de recuperação de informação. Primeiramente,

descrevemos os participantes da área da saúde envolvidos nas atividades de avaliação e os cenários que expressam consultas com necessidades do ponto de visa clínico.

Em seguida, apresentamos como as anotações de ilocução foram incluídas no sistema, bem como o procedimento conduzido para a construção da base de referência para cada cenário com o objetivo de avaliar a proposta de solução. Após isso, analisamos os benefícios da Busca Pragmática segundo medidas de avaliação padrão da área. A Busca Pragmática foi comparada com técnicas de Busca Sintática utilizando o *software Solr* e com a Busca Expandida usando a técnica de expansão de termos no domínio por meio do UTS.

Os testes experimentais demonstraram a eficiência global da técnica de recuperação desenvolvida nesta dissertação de mestrado e seus benefícios para o domínio médico. A Busca Pragmática obteve a melhor *Medida-F* nos cenários avaliados, com avanço expressivo na precisão dos resultados quando comparados à Busca Sintática e Busca Expandida.

Por fim, os resultados foram discutidos e as limitações do método e do experimento apresentadas. Apontamos ainda desafios a serem abordados em investigações futuras que podem aprimorar a eficácia, a qualidade dos resultados revelados e a escalabilidade da proposta. O próximo capítulo finaliza esta dissertação sumarizando as contribuições e próximos passos da pesquisa.

Capítulo 7

Conclusão

A recuperação de informação em sistemas de *software* médicos pode beneficiar os pacientes de diversas maneiras, incluindo cenários ligados a tratamentos, gestão e pesquisas médicas. Esses sistemas armazenam um número crescente de PEPs, que são relevantes por conter dados sobre o histórico de saúde dos pacientes. Entretanto, textos pouco estruturados descritos em linguagem natural dificultam a recuperação desses dados.

Embora a literatura apresente diversos modelos e mecanismos de busca aplicados ao contexto de sistemas médicos, as propostas atuais estão limitadas a lidar principalmente com o processamento sintático e semântico da informação. Esta dissertação de mestrado investigou um método original de recuperação de informação em repositórios PEPs. A proposta adotou conceitos e métodos da Teoria dos Atos da Fala e da Semiótica Organizacional para estruturar e considerar categorias de intenções no processamento de consultas.

7.1. Contribuições da Pesquisa

Esta dissertação contribuiu para o avanço de conhecimento em recuperação de informação com o foco na área de saúde. A contribuição central do trabalho foi entender como os aspectos pragmáticos podem beneficiar a recuperação de informação ao considerar as intenções declaradas pelos usuários. Esta pesquisa atingiu diversas contribuições específicas conforme destacadas a seguir.

Estudo sobre intenções expressas em PEPs. Nossa primeira análise examinou como as intenções se manifestaram em um conjunto real de PEPs. Resultados apontaram como as principais classes de ilocução estão presentes nos conteúdos examinados. Este estudo foi relevante para definirmos a ocorrência de termos que são relevantes para caracterizar as ilocuções. Por exemplo, revelou os valores típicos que podem ser associados a uma dimensão como um parâmetro da consulta no mecanismo de busca.

Método de Recuperação de Informação Médica com Base em Ilocuções. Esta pesquisa definiu meios para considerar tipos de intenções em um motor de busca. Propomos um método que organiza os principais elementos envolvidos na solução, incluindo as etapas de anotação semântica e pragmática. O método permitiu explorar o conhecimento do domínio por meio de Sistemas de Organização do Conhecimento, tais como o UMLS. Elaboramos procedimentos para efetuar expansão de consultas por meio desses artefatos e combinamos a expansão semântica dos termos com filtros mais precisos relacionados às ilocuções.

Algoritmo de seleção e filtro de PEPs. No núcleo da proposta concebemos e desenvolvemos o algoritmo *PraSA (Pragmatic Search Algorithm)* que formaliza as etapas da busca pragmática. O algoritmo permite recuperar documentos relacionados às palavras-chave definidas e expandir termos da consulta. Ele filtra os resultados de busca dando prioridade às ocorrências com base nas ilocuções, seguindo critérios em relação aos parâmetros de consulta e anotações disponíveis. Demonstramos a execução do algoritmo em um cenário que permitiu observar suas características e comportamento.

O Sistema SiRBI. A proposta conceitual foi implementada em um protótipo executável denominado Sistema de Recuperação com Base em Intenções (SiRBI). O protótipo permite aos usuários expressarem as consultas e obterem os resultados de busca com base no PraSA. Desenvolvemos uma arquitetura que inclui os componentes necessários ao funcionamento da abordagem. Apresentamos, ainda, como os PEPs são tratados através da análise de termos dos seus conteúdos. Para tanto foi adotada a plataforma de indexação *Apache Lucene Solr*, que é tradicionalmente usada em sistemas de recuperação de informação. O trabalho herda e reusa a ordenação dos resultados amplamente estudada nessa plataforma.

Validação Experimental da Proposta. O sistema SiRBI permitiu conduzirmos avaliações experimentais com o objetivo de validar os conceitos da abordagem. Profissionais de saúde foram envolvidos na avaliação para garantir rigor técnico e consistência nos resultados. Eles definiram cenários de recuperação relevantes para a área e usaram o sistema para gerar uma base de referência ao classificar PEPs adequados para cada cenário. Os experimentos demonstraram que o uso das intenções resultou na melhoria de qualidade nos PEPs recuperados para os cenários estudados. Esta

investigação evidenciou avanços constatados por meio de medidas padrão, ao considerar intenções como filtro pragmático com técnicas de recuperação sintática e semântica. Os cenários definidos permitiram examinar diferentes níveis de complexidade na recuperação.

Publicações Científicas. O trabalho desenvolvido nesta dissertação de mestrado permitiu a publicação de dois artigos completos em conferências internacionais. Um artigo foi publicado na 17th *International Conference on Informatics and Semiotics in Organisations* (ICISO 2016). Recebeu prêmio de melhor artigo completo na 25th IEEE *International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises* (WETICE 2016) (*cf.* Apêndice I).

7.2. Trabalhos Futuros

Esta pesquisa apontou diversas soluções e alcançou os objetivos e respostas para as questões de pesquisa levantadas. Contudo, como um trabalho pioneiro, a abordagem possui limitações e oferece perspectivas relevantes para trabalhos futuros em termos de validação experimental, aprimoramento e extensões.

Para avaliar de maneira mais consistente a abordagem proposta, sugerimos explorar bases de teste disponíveis a partir de eventos dedicados a avaliações em recuperação de informação, *e.g.*, *Text REtrieval Conference* (TREC). O grande desafio neste contexto é adequar os cenários e necessidades de anotação inerentes da proposta.

Para tornar a proposta escalável é necessário conceber técnicas semi-automáticas de anotação de ilocuções. Portanto, objetiva-se construir um mecanismo de anotação para classificação e reconhecimento das ilocuções contidas nos PEPs. Ainda que as anotações sobre as intenções sejam realizadas manualmente, nosso método mostrou ser factível para pequenas bases. Nesta perspectiva, a pesquisa enfrentará desafios relacionados ao processamento de linguagem natural e pode explorar terminologias no domínio como meio de auxiliar no processo. As anotações manuais efetuadas apontaram certos padrões que podem ser reusados para a definição da técnica de anotação.

Uma extensão desta pesquisa refere-se ao tratamento dos valores das dimensões por meio de terminologias do domínio. Por exemplo, detectar automaticamente que “período noturno” é equivalente a “noite”, “há mais de 7 dias” equivale a “mais de uma

semana”, *etc*. Adicionalmente, investigar no *ranking* o quão distante o conteúdo está do parâmetro. Por exemplo, se na consulta foi especificado há menos de 4 semanas, um PEP que contenha um fragmento de texto “1 semana e meia” deve vir primeiro que outro que descreve algo relacionado a “1 semana”.

Estudos adicionais também serão desenvolvidos no refinamento da fase de ordenação dos resultados. Pretendemos avaliar diversas configurações e medir a influência de forma mais precisa das ilocuções.

Outro aspecto a ser abordado nos próximos passos dessa pesquisa é a melhoria de desempenho relativa ao tempo de execução em função do número de registros e parâmetros de consulta. Para tanto, devem ser pesquisados o uso de arquiteturas escaláveis e a otimização da implementação do protótipo e algoritmo. Por fim, objetivamos implantar a solução em um ambiente hospitalar para uso em produção em longo prazo, e assim coletar resultados e buscar desafios futuros.

Referências

Asai, H. & Yamana, H. (2014). Intelligent Ink Annotation Framework that uses User's Intention in Electronic Document Annotation. *ITS '14 Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, pp. 333-338.

Austin, J.L. (1962).*How to Do Things with Words*, Oxford University Press, Oxford, England.

Austin, J.L. (1976). *How To Do Things With Words*, Second Edition , Oxford University Press.

Austin, J. L. (1990). Quando dizer é fazer: palavras e ação. *Tradução de Danilo Marcondes de Souza Filho. Porto Alegre: Artes Médicas*, pp 136.

Bhatt, M., Rahayu, W. & Soni, S.P. (2009). Ontology driven semantic profiling and retrieval in medical information systems. *Semantic Web challenge 2008*, vol. 7, Issue 4, pp. 317-331.

Bo, C., Yang-Mei, L.(2015). Design and Development of Semantic-based Search Engine Model. *Intelligent Computation Technology and Automation (ICICTA), 2014 7th International Conference on*, pp 145-148.

Bonacin, R., Reis, J.C., Hornung, H. & Baranauskas, M.C. (2013). An ontological model for supporting intention-based information sharing on collaborative problem solving. *International Journal of Collaborative Enterprise*, vol. 3, pp. 130-150.

Bonacin, R. (2004). Um modelo de desenvolvimento de sistemas para suporte a cooperação fundamentado em design participativo e semiótica organizacional (Tese de Doutorado). Disponível em*: Biblioteca Digital da UNICAMP:* http://www.bibliotecadigital.unicamp.br/document/?code=vtls000319294&fd=y. Acessado em 14 de maio de 2015.

Cenan, C. (2008). A proposed architecture and ontology for a software system for managing patient's health record. *IEEE International Conference on Automation Quality and Testing, Robotics, 2008. AQTR 2008*, vol. 3, pp. 123-127.

Chahal, P., Singh, M. & Kumar S. (2013). Ranking of *Web* Documents using Semantic Similarity. *International Conference on Information Systems and Computer Networks*, pp. 145-150.

Chawla, S., Bedi, P. (2008). Query Expansion using Information Scent. 2008 *International Symposium on Information Technology*, vol 3, pp. 1-8.

Chen, T., Chung, Y. & Lin, F. (2012). A Study on Agent-Based Secure Scheme for Electronic Medical Record System. *Journal of Medical Systems*, vol 36, Issue 3 , pp. 1345-1357.

Cogley, J., Stokes, N. & Carthy, J. (2013). Exploring the effectiveness of Medical Entity Recognition for Clinical Information Retrieval. *DTMBIO '13 Proceedings of the 7th international workshop on Data and text mining in biomedical informatics*, pp. 3-4.

Conselho Federal de Medicina. *Resolução CFM nº 1.638/2002 - (Publicada no D.O.U. de 9 de agosto de 2002, Seção I, p.184-5).* Disponível em: http://www.portalmedico.org.br/resolucoes/cfm/2002/1638_2002.htm. Acesso em 20 de outubro 2015.

Dias, T. D. & Santos, N. (2001). *Web* Semântica: Conceitos Básicos e Tecnologias Associadas (Tutorial). *Monografia (Bacharelado em Ciência da Computação) - Instituto de Matemática e Estatística - Universidade do Estado do Rio de Janeiro, Rio de Janeiro*, 14 p.

Diem, L. H., Chavallet, J., Thuy, D. T. B. (2007). Thesaurus-based query and document expansion in conceptual indexing with UMLS. *Research, Innovation and Vision for the Future, 2007 IEEE International Conference on (2008)*, pp-242-246.

Dos Reis, J. C., Bonacin, R, Baranauskas, M. C. C. (2014). Addressing universal access in social networks: an inclusive search mechanism. *In International Journal of Universal Access in the Information Society (UAIS)*. vol. 13, Issue 2, pp. 125-145.

Dong, H., Hussain, F.K. & Chang, E. (2008). A survey in semantic search technologies. *2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008. DEST 2008*, pp. 403 – 408.

Forcher, B., Berghofer, T.R., Agne, S. & Dengel A. (2014). Intuitive justifications of medical semantic search results. *Engineering Applications of Artificial Intelligence,* vol. 30, pp. 1–17.

Guha, R. McCool, R. & Miller, E. (2003). Semantic Search. *Proceedings of the 12th international conference on World Wide Web*, pp. 700-709.

Gupta, V. Garg, N. & Gupta, T. (2012). Search Bot: Search Intention Based Filtering Using Decision Tree Based Technique. *2012 Third International Conference on, Intelligent Systems, Modelling and Simulation (ISMS),* pp. 49-54.

Gurulingappa, H., Muller, B., Apitius, M.H. & Fluck, J. (2011). A Semantic Platform for Information Retrieval from E-Health Records. *In The Twentieth Text RETrieval Conference (TREC 2011) Proceedings, Gaithersburg, Maryland, USA.*

Hambury, A. (2012). Medical Information Retrivial- An Instance of Domain-Specific. *Search. S SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1191-1192.

Hannan, B., Zhang, X. & Sethares, K. (2014). iHANDs: Intelligent Health Advising and Decision-Support *Agent. 2014 IEEE/WIC/ACM International Joint Conferences on, Web Intelligence (WI) and Intelligent Agent Technologies (IAT),* vol 3, pp. 924-301.

Heard, S. & Beale, T. (2007). *openEHR* Architecture: Architecture Overview. *openEHR Specification Program, openEHR Foundation,Revision: 1.1.2*, pp.78. Disponível em: https://github.com/openEHR/specifications/blob/master/publishing/architecture/overview.pdf?raw=true. Acesso em 09 de agosto de 2015.

Hildebrand, M.; Ossenbruggen J., & Hardman, L.. (2007): An analysis of search-based user interaction on the semantic web. *Report, CWI, Amsterdam, Holland,* pp. 18.

Hwang, M., Kim, P. & Choi, D. (2011). Information Retrieval Techniques to Grasp User Intention in Pervasive Computing Environment. *2011 Fifth International Conference on, Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS),* pp. 186-191.

Ilari, R. “Introdução à semântica: brincando com a gramática”. *5. ed. São Paulo: Contexto, 2003*, pp 206 .

Jaladi, V., Borujerdi, M., R., M. (2008). The Effect of Using Domain Specific Ontologies in Query Expansion in Medical Field. *Innovations in Information Technology, 2008. IIT 2008. International Conference on*, pp. 277-281.

Jenice, A.R. & Kurian, M. (2012). A Semantic *Web*: Intelligence in Information Retrieval. *Department of Computer Science and Engineering, Karunya University, India*. pp. 203 – 206.

Kassim, J.M. & Rahmany, M. (2009). Introduction to Semantic Search Engine. *ICEEI'09. International Conference on Electrical Engineering and Informatics, 2009,* vol. 2, pp. 380-386.

Laforest, F. & Tchounikine, A. (1999). Indexing Semi-Structured Documents for Context-based Information Retrieval in a Medical Information System. *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, pp.593-597.

Liaw, S.T., Taggart J., Yu, H., Lusignan, S., Kusiemsky, C. & Hayen, A. (2014). Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers. *Journal of Biomedical Informatics*, vol. 52, Issue C, pp. 364-372.

Liu, K. (2000). Semiotics in Information Systems Engineering. Cambridge University Press. Cambridge.

Montes-y-Gomez, M., Gelbukh A.F. & Lopez-Lopez, A. (1999). Document Title Patterns in Information Retrieval. *Springer Berlin Heidelberg*, pp. 372-375.

Mendoza, M. & Baeza-Yates, R. (2008). A *Web* Search Analysis Considering the Intention behind Queries. *Web Conference, 2008. LA-WEB '08, Latin American*, pp. 66-74.

Morris, C. (1937). Logical positivism, pragmatism, and scientific empiricism (Philosophy in America). Hermann et Cie; 1st edit edition.

Muller, H., Zhou, X., Depeursinge, A. , Pitkanen, M., Iavindrasana, J. & Geissbuhler, A. (2007). Medical Visual Information Retrieval: State of the Art and Challenges Ahead. *2007 IEEE International Conference on, Multimedia and Expo,* pp. 683 – 686.

Noor, S. & Martinez, K. (2009). Using social data as context for making recommendations: an ontology based approach. *CIAO '09 Proceedings of the 1st Workshop on Context, Information and Ontologies. Article No 7*.

Peirce, C. S. (1931-1958). Collected Papers. Cambridge: Harvard University Press.

Popescu, M. (2010). An Ontological Fuzzy Smith-Waterman with Applications to Patient Retrieval in Electronic Medical Records. *2010 IEEE International Conference on, Fuzzy Systems (FUZZ),* pp. 1 – 6.

Pruski, C. & Wisniewski, F. (2012). Efficient Medical Information Retrieval in Encrypted Electronic Health Records. *Quality of Life through Quality of Information J. Mantas et al. (Eds.) IOS Press*, pp. 225-229.

Rey, D.P., Castellannos, A.J., Remessal, M.G., Crespo, J. & Maojo, V. (2012). CDAPubMed: a browser extension to retrieve EHR-based biomedical literature. *BMC Medical Informatics and Decision Making*, pp. 1-10.

Santaella, L.(2004). 'Comunicação e semiótica'. *São Paulo: Hacker*.

Searle, J.R. (1969) 'Speech acts', *Cambridge University Press, Language in Society,* vol.5, Issue 1, pp.1–23.

Searle, J.R. (1976) 'A classification of illocutionary acts', Language in Society, vol. 5, Issue 1, pp.1–23.

Sharef, N.M., Madzin, H. (2012). IMS: An improved medical retrieval model via medical-context aware query expansion and comprehensive ranking. *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*. pp-214-218

Tang, X., Liu, K., Cui, J., Wen, F. & Wang, X. (2012). IntentSearch: Capturing User Intention. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, Issue 7, pp. 1342-1353.

Tawfik, A.A., Kochendorfer, K.M., Saparova, D., Al Ghenaimi, S. & More, J.L. (2011). Using semantic search to reduce cognitive load in an electronic health record. *2011 13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom),* pp. 181 – 184.

Totelin, L.M.V.(2006). Hippocratic Recipes: Oral and Written Transmission of Pharmacological Knowledge in Fifth- And Fourth-Century Greece. *Editora**:** Brill*.

W3C Brasil 2012.World Wide *Web* Consortium. Disponível em http://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211. Acessado em agosto de 2015.

Weed, L. (1968). Medical Record that Guide and teache. *The New England Journal of Medicine,* vol 238, pp 593 – 600.

Zhang, J., Huang, J.X., Guo, J. & Xu, W. (2013). Promoting Electronic Health Record Search through A Time-aware Approach. *2013 IEEE International Conference on, Bioinformatics and Biomedicine (BIBM),* pp. 593-596.

Zinglé, H. (2006). Modelling Knowledge with ZDoc for the Purposes of Information Retrieval. *Springer Berlin Heidelberg*, pp. 1053-1058.

Apêndice I — Artigos Publicados

1. DOS REIS, J. C.; BONACIN, R.; PERCIANI, E. M. 2016. ***Intention-based Information Retrieval of Electronic Health Records.*** Anais da 25th IEEE *International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises* (WETICE 2016), Paris, França. pp. 217–222, DOI:10.1109/WETICE.2016.56. [**prêmio de melhor artigo completo**]

   Abstract *— The large volume of information stored in electronical health records is very valuable in the medical field, e.g., for clinical research and administrative purposes. However, health care professionals still face difficulties to recover and select relevant data. Although literature has investigated the influence of lexical, syntactical and semantic parameters in information retrieval techniques, few studies explore intentions as explicit users' actions in information recovery. This article studies means of considering intentions in search engines. We define a method with an original algorithm to properly rank search results from a set of electronical health records. This investigation relies on well-established theories to categorize several types of intentions. The proposed technique is based on knowledge representation models for annotating meanings and intentions in medical records. Achieved results are illustrated in scenarios based on real medical cases.*

   Keywords *— information retrieval; intentions; illocutions; pragmatics; semantic search; eletronic health records.*

2. DOS REIS, J. C.; BONACIN, R.; PERCIANI, E. M.; BARANAUSKAS, M.C.C. 2016. ***Analysis and Representation of Illocutions from Electronic Health Records.*** Anais 17th *International Conference on Informatics and Semiotics in Organisations* (ICISO 2016), Campinas, Brazil. pp. 209–218, DOI: 10.1007/978-3-319-42102-5_24.

   Abstract *— Electronic Health Records (EHRs) describe various patients" data, including medical history, diagnoses and treatments. Computer-interpretable representation of meanings and intentions of EHRs content might play a major role for decision making, as well as for medical system integration and information recovery. However, there is a lack of suitable representation models to specify the*

relations between semantic models and illocutions, which reflect the intentions of medical content producers. In this article, we propose an analysis to understand how illocutions are expressed in EHRs. We aim to identify the domain-specific terms to convey the different dimensions in which illocutions are classified. Furthermore, this research develops a model, based on ontology description languages, to encode and instantiate the illocutions in the medical domain. Obtained results point out that some illocution types and associated terms are predominant in the analyzed content. They highlight the potentially of our model to explore illocutions in several computing tasks. ***Keywords*** *— Intentions; Illocutions; Pragmatics; Ontologies; Semantic Web; Pragmatic Web; Knowledge Representation; EHRs; Medical data.*