UNIVERSIDADE DE SÃO PAULO
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Computação e Matemática

Manipulação e Relacionamento de Informação

ALESSANDRA ALANIZ MACEDO

Ribeirão Preto
2016

UNIVERSIDADE DE SÃO PAULO
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Computação e Matemática

Manipulação e Relacionamento de Informação

ALESSANDRA ALANIZ MACEDO

Documento submetido à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da UNIVERSIDADE DE SÃO PAULO como parte dos requisitos para a obtenção do título de Livre-Docente na Área de Ciências de Computação, especialidade: Processamento, Manipulação e Recuperação de Informação.

Ribeirão Preto, maio de 2016.

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE DOCUMENTO, POR MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Macedo, A. A.

Manipulação e Relacionamento de Informação/ Alessandra Alaniz Macedo – Ribeirão Preto/SP, 2016.

114p .: il.

Documento (Livre-Docente. Área de Ciências de Computação, especialidade: Processamento, Manipulação e Recuperação de Informação) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da UNIVERSIDADE DE SÃO PAULO.

1. Processamento Textual de Informação 2. Fusão de Informações 3. Ligação

DEDICATÓRIA

Dedico este trabalho ao Antonio, aos meus filhos e aos amigos do DCM.

*Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?*

Choruses from *The Rock* by T.S. Eliot, 1934

Agradecimentos

A Deus pela presença na minha vida.

À Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), a CAPES e ao CNPq, que incentivam, financiam e dão suporte à pesquisa neste país.

Aos professores-colegas de profissão, que compartilham a paixão pelo ensino e pela pesquisa e a esperança na transformação do mundo, utilizando essas ferramentas, principalmente ao Prof. Baranauskas com quem tenho colaborado em trabalhos e dividido angústias.

Aos pesquisadores-colaboradores do Brasil e do exterior, que apoiam e discutem os trabalhos vislumbrados por mim.

Ao pessoal, principalmente aos meus alunos, do laboratório 621 do DCM, que ajudou a traduzir ideias e pensamentos em trabalhos científicos.

Aos funcionários e prestadores de serviço do DCM pelo suporte de infraestrutura.

À minha família e amigos pelo apoio, carinho e amor eternos.

Aos meus sogros e familiares distantes que, mesmo a milhares de quilômetros, sempre me incentivam e enviam muita energia positiva.

Ao meu marido, José Antonio, e aos meus filhos, Helena e Santiago, pelo apoio e pela paciência inestimável!!

Sumário

Lista de Figuras	p. iv
Normas e convenções	p. vi
Resumo	p. vii
Abstract	p. viii
1 Introdução	p. 1
2 Manipulação e Relacionamento de Informação Textual	p. 5
2.1 Criação Automática de Ligações	p. 6
2.1.1 <i>Framework</i> para Criação Automática de Ligações	p. 7
2.2 Outros Mecanismos de Criação Automática de Ligações	p. 11
2.2.1 Criação Automática de Ligações entre Informações Multilíngues . . .	p. 11
2.2.2 Criação Automática de Ligações a partir do Suporte de RSS	p. 13
2.2.3 Criação Automática de Ligações a partir de <i>Clustering</i> e da Catego- rização de Informações	p. 14
2.2.4 Criação Automática de Relacionamentos em Engenharia de Software	p. 15
2.3 Busca Guiada de Patentes	p. 17
2.4 Considerações Finais	p. 20
3 Manipulação e Relacionamento de Informação Multimídia e Multimodal	p. 21
3.1 Relacionamento de Informação Multimodal	p. 22
3.1.1 Contexto Educacional	p. 23

3.1.2	Contexto de Saúde e de Medicina	p. 25
3.1.3	Contexto de Casas Inteligentes	p. 27
3.2	Relacionamento de Informações de <i>Optical Character Recognition</i> (OCR) . .	p. 28
3.3	Relacionamento de Informações Textuais e de Imagens	p. 30
3.3.1	Contexto de <i>Computer-Aided Diagnosis</i> (CAD)	p. 30
3.3.2	Contexto de <i>Picture Archiving and Communication System</i> (PACS) .	p. 34
3.4	Relacionamento de Informações Textuais e de Vídeo	p. 35
3.5	Considerações Finais	p. 37
4	Manipulação, Relacionamento e Classificação de Informação Biomédica na Prevenção a Doenças	p. 38
4.1	<i>Automatic Surveillance Level</i> (Automatic-SL)	p. 39
4.2	<i>Chronic Illness Surveillance System</i> (CISS)	p. 43
4.3	<i>Health Surveillance Systems Framework</i> (HSSF)	p. 47
4.3.1	Infraestrutura de Software	p. 47
4.3.2	Diagrama de Classes e Pontos de Reúso	p. 50
4.3.2.1	Identificação de Pontos Flexíveis e de Pontos Fixos	p. 50
4.3.2.2	Diagrama de Classes	p. 51
4.3.3	<i>New Chronic Illness Surveillance System</i> (CISS+)	p. 53
4.3.4	<i>Chronic Illness Surveillance System by Semantic Web</i> (CISS-SW) . .	p. 55
4.4	<i>Question Answering Surveillance Framework</i> (QASF)	p. 56
4.5	Sumarização de Artigos Científicos em Interações Gênicas (SARI)	p. 59
4.6	Assistência de Gestantes para Vigilância de Fatores de Risco	p. 62
4.7	Considerações Finais	p. 63
5	Manipulação e Relacionamento de Informações com Redes Complexas	p. 65
5.1	Conceitos Gerais	p. 66
5.2	Redes Complexas para Investigação de Relacionamento de Informações . . .	p. 70

5.2.1	Redes Complexas Heterogêneas	p. 71
5.3	Redes Complexas Heterogêneas para Informações Biomédicas	p. 73
5.3.1	Materiais e Métodos	p. 74
5.3.2	Metodologia	p. 77
5.4	Considerações Finais	p. 81
6	Conclusão	p. 82
6.1	Dificuldades Encontradas	p. 84
	Referências	p. 85

Lista de Figuras

2.1	Diagrama de Classes do <i>framework</i> CARE para apoio à criação automática de ligações entre repositórios homogêneos de informações [1].	p. 9
2.2	Medida F, precisão e revocação para relacionamentos entre casos clínicos no domínio de Oftamologia a partir da Scielo [2].	p. 13
3.1	Infraestrutura da CALiSP (<i>Infrastructure for Capturing, Accessing, Linking, Storing and Presenting information</i>) [3].	p. 24
3.2	Arquitetura ArcaMed [4].	p. 26
3.3	Visão geral do LinkDI para geração de relacionamentos considerando imagens de documentos [5] [6].	p. 29
3.4	Classes de Software do FREDs. Em cinza, os componentes propostos em [7] e em branco os desenvolvidos em [8].	p. 33
3.5	Método proposto para um SADM suportado pelo FREDs [9].	p. 34
3.6	Avaliação Geral [10].	p. 37
4.1	(1) Arquitetura do sistema Automatic-SL organizada em três camadas: a) Camada de Apresentação, b) Camada de Classificação, e c) Camada de Armazenamento. A Camada de Classificação tem cinco módulos de classificação: KNN-SL, ANN-SL, RF-SL, DT-SL e vote-SL e um módulo de linguística [11].	p. 42
4.2	Módulos de processamento e de armazenamento da infraestrutura CISS. Uma <i>Graphical User Interface</i> (GUI) pode ser apresentada aos pediatras para ilustrar a recuperação de artigos científicos em inglês após a consulta implícita considerando um registro médico em português [12].	p. 46
4.3	Infraestrutura do HSSF composta por módulos de processamento e de armazenamento [13].	p. 49
4.4	Diagrama de Classes [14].	p. 52

4.5	Arquitetura geral de um sistema de <i>Question Answering</i> junto aos submódulos do QASF [15].	p. 56
4.6	Diagrama de classes do HSSF com o QASF.	p. 59
5.1	Entidades e Relacionamentos iniciais da Rede Complexa Heterogênea de Informação Biomédica.	p. 76

Normas e convenções

Este documento foi preparado com o formatador de textos $\text{\LaTeX}$. O sistema de citações de referências bibliográficas utiliza a classe *ieeetr* do $\text{\BIBTeX}$, que segue as recomendações do IEEE (*Institute of Electrical and Electronics Engineers*) para publicação em periódicos da instituição.

A formatação da capa, folha de rosto, folha de aprovação, resumo e abstract segue as “diretrizes para apresentação de dissertações e teses da USP”, disponível em <http://www.teses.usp.br>.

A formatação de sumário, lista de figuras e tabelas, lista de abreviaturas e siglas, espaçamento entre linhas, numeração de páginas e cabeçalhos de páginas segue a norma ABNT NBR 14724 para “Apresentação de trabalhos acadêmicos”.

A formatação de títulos e capítulos de seções segue a norma ABNT NBR 6024 para “Numeração progressiva das seções de um documento”.

Todas as formatações que seguem a norma ABNT foram geradas automaticamente utilizando as macros da classe *abntex* disponíveis em <http://abntex.codigolivres.org.br/>.

Resumo

MACEDO, A. A.. **Manipulação e Relacionamento de Informação.** Documento (Livre-Docente) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2016.

Múltiplas fontes de informação, maneiras diferentes de escrita, estilos distintos de apresentação e os diferentes anseios do usuário (curiosidade, tomada de decisão, análise, conhecimento e etc) pela informação sugerem a necessidade de obtenção de diferentes relatos similares a um mesmo assunto. Por exemplo na área de saúde, profissionais reúnem informações, as quais variam em tipos de mídia (texto, imagem e sinais), estrutura (prontuário eletrônico, laudos, imagens de raio-X, tomografias, análises clínicas de exames etc) e conteúdo, para a tomada de decisões. Nesse e outros exemplos, as informações advindas em diferentes mídias são informações complementares e podem agregar valor semântico para o entendimento do assunto em questão. Porém para que a leitura de diferentes relatos de informações seja possível, leitores precisam definir mecanismos de buscas, procurar, selecionar, ler e analisar as informações fornecidas a eles. Essas atividades, além de exigirem grande investimento de tempo, sobrecarregam cognitivamente os leitores que dedicam atenção à informação lida, à trilha percorrida e aos mecanismos preparados para essa pesquisa. Este texto apresenta grande parte do trabalho desenvolvido pela pesquisadora em sua vida acadêmica. O texto foi sistematizado em quatro partes da seguinte maneira: a primeira parte aborda os problemas de relacionamento de informação textual em repositórios Web; a segunda extrapola o formato da informação manipulada e o ambiente de captura da informação para definir os relacionamentos entre informações multimídia e multimodal; a terceira etapa apresenta mecanismos de relacionamento de informações para a manipulação de informações biomédicas; a quarta etapa, em execução, está buscando representar os conceitos e resultados dos trabalhos em grafos semânticos para posterior análises das propriedades de redes complexas heterogêneas. Além desta divisão explícita de tema, a mesma divisão segue uma linha temporal, quase exata, referente ao desenvolvimento das pesquisas desde o doutoramento desta pesquisadora até os dias atuais. Em todas as partes do texto, pretende-se ilustrar o domínio do assunto abordado, bem como resumir as contribuições da pesquisadora nas áreas de conhecimento envolvidas, deixando citações para o detalhamento quando necessário. Em sua vida acadêmica, esta docente encontrou situações laboriosas que foram superadas com reflexões, mudanças e enfrentamento de novos desafios. As pesquisas e os resultados apresentados neste texto envolvem temas atuais, multidisciplinares e de grande interesse na comunidade científica.

Palavras-chave: extração de informação, recuperação de informação, hiperligações.

Abstract

MACEDO, A. A.. **Information Manipulation and Relationship**. Document (Livre-Docente) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2016.

Multiple information sources, different writing styles, forms of presentation and aspirations by the user (curiosity, decision making, analysis, knowledge, etc.) in relation to information suggest the need to obtain different but similar reports on the same subject. For example, to make decisions health care professionals gather information that vary in terms of media type (text, image and signal), structure (electronic medical records, reports, X-ray images, CT scans, clinical analysis tests, etc.) and content. In this and other examples, information resulting from different media is complementary information and can add semantic value to the understanding of the subject in question. However, for the reading of different information reports to be possible, readers must define search mechanisms, search, select, read and analyze the information provided to them. In addition to demanding a lot of time, such activities cognitively overwhelm readers, who have to devote their attention to the information read, to the path followed, and to mechanisms prepared for this search. This text presents much of the work conducted by the researcher throughout her academic life. The text is structured in four parts: the first deals with problems related to textual information relationship in Web repositories; the second extrapolates the format of manipulated information and the information capture environment to define the relationships between media and multimodal information; the third presents information relationship mechanisms for the manipulation of biomedical information; and the fourth, currently under development, is an attempt to represent the concepts and results of the work in semantic graphs for further analysis of the properties of complex heterogeneous networks. In addition to such explicit thematic subdivision, this division follows a nearly exact timeline in relation to the works conducted by the researcher from her doctoral studies to the present day. All parts of the text are intended to illustrate the domain of the subject matter, as well as summarize the contributions by the researcher in the areas of knowledge involved. Quotations are left for further detailing when necessary. Along her academic journey, the researcher has found laborious situations that were overcome through reflections, changes and by facing new challenges. The research and results presented involve current and multidisciplinary issues of great interest to the scientific community.

Keywords: information extraction, information retrieval, hyperlink.

1 *Introdução*

Os diferentes anseios dos usuários (curiosidade, tomada de decisão, análise, conhecimento e etc) pela informação, suportados por múltiplas fontes de informação, diferentes escrita, estilos distintos de apresentação sugerem a necessidade de obtenção de diferentes relatos de informações similares sobre um mesmo assunto. No contexto de pesquisas científicas, por exemplo, a identificação de trabalhos relacionados é condição essencial para o desenvolvimento de um trabalho. Na área de saúde, em discussões clínicas, os profissionais reúnem informações, as quais variam em tipos de mídia (texto, imagem e sinais), estrutura (prontuário eletrônico, laudos, imagens de raio-X, tomografias, análises clínicas de exames etc) e conteúdo, para a tomada de decisões. No dia a dia, uma fotografia de um jogo de futebol pode ser associada ao evento correspondente apenas quando a data e o local são conhecidos. Provedores de informação na Internet normalmente organizam o conteúdo a ser disponibilizado como documentos Web, os quais contêm informações textuais, imagens, vídeos e ligações hipertexto ou hiperligações¹. A Wikipedia é uma coleção de informações disponibilizada na Web, e essa coleção combina diferentes tipos de mídias (textos ilustrados com imagens, gráficos, logotipos e pinturas) para apresentar seus conteúdos aos leitores. Nos casos das reuniões clínicas e da Wikipedia, as informações advindas de diferentes mídias são informações complementares e podem agregar valor semântico para o entendimento do assunto em questão. Para que a leitura de diferentes relatos de informações complementares e/ou relacionadas, descritas em texto ou em diferentes mídias, seja possível, leitores precisam definir mecanismos de buscas, procurar, selecionar, ler e analisar as informações a eles fornecidas. Essas atividades, além de exigirem grande investimento de tempo, sobrecarregam cognitivamente os leitores, que dedicam atenção à informação lida, à trilha percorrida, e aos mecanismos preparados para essa pesquisa. Os exemplos apresentados ilustram alguns cenários que demandam a intenção do usuário em buscar informações correlatas. Todavia, existem situações em que a sugestão de informação relacionada ou complementar pode partir do sistema computacional, após a permissão do usuário.

---

¹ A base da tecnologia de hipertexto é a rede de informações que possui interconexões, as quais devem estar facilmente acessíveis para os usuários. Essa rede de informações, com os “nós” são as unidades de informação e as “ligações” (ou *links*) as interconexões entre eles, compõe um hiperdocumento.

As diversidades existentes em termos de estruturação do conjunto de documentos e de estruturação do conteúdo de cada documento geram desafios cognitivos para leitores e desafios científicos para pesquisadores, desde remotas gerações de cientistas, principalmente em áreas relacionadas à manipulação e ao processamento de informação. Motivado pela aplicação da ciência em assuntos bélicos em 1945, Vannevar Bush publicou o artigo “As We May Think” para descrever um dispositivo no qual um indivíduo armazenava todos os seus livros, os registros e as comunicações, e que era mecanizado para que tudo isso fosse consultado com a máxima velocidade e flexibilidade (uma memória aumentada) [16]. Bush contemplava a indexação associativa, isto é, uma provisão pela qual qualquer item pode ser ativado para selecionar imediatamente e automaticamente um outro. O mesmo Bush afirmava que “o processo de ligar dois itens é o que é importante”. Motivado pelo artigo de Bush, Douglas Engelbart imaginava o uso de ferramentas baseadas em computador para aumentar o intelecto humano, e melhorar a capacidade global do ser humano para resolver os problemas. Em seu trabalho no Instituto Bootstrap, Engelbart cunhou o termo “QI coletivo” para descrever como um grupo pode alavancar sua memória coletiva, percepção, planejamento, raciocínio, previsão e experiência em conhecimento aplicável para resolver os problemas dos usuários [17]. Em 1989, Tim Berners-Lee, utilizando conceitos de Bush, propõe a criação da Web, a qual tem constantemente evoluído por meio da criação de protocolos e de padrões que buscam aprimorar a organização, a recuperação, a manipulação, e a associação de informações para garantir o crescimento da Web a longo prazo [18].

No moderno contexto de computação ubíqua ou mesmo dos ambientes “inteligentes”, pesquisadores de diversas áreas sugerem a agregação, de forma contínua e não-intrusiva, de serviços em diferentes ambientes do cotidiano para tornar esses ambientes “inteligentes”, ou pelo menos, mais cientes das informações de contexto. Nesse sentido, além da ciência, dados e informações multimodais capturados por diferentes dispositivos podem ser relacionados para prover inúmeros benefícios aos usuários, como a geração de conhecimento e de memória, inclusive, a oportunidade de reúso de informações e de experiências. A pesquisa translacional em saúde e medicina se beneficia diretamente da associação de informações, sejam elas representadas por documentos Web, por informações multimodais, ou mesmo por registros médicos de pacientes, os quais podem ser relacionados a informações científicas, e podem beneficiar o paciente. Em qualquer domínio, múltiplos relatos de um problema permitem potencialmente a redução da *gap* semântico entre as informações, além da redução da discrepância entre as intenções do usuário e o resultado obtido.

A sinergia entre as informações de diferentes tipos, e em grande volume, pode ser representada de diversas maneiras. Na atualidade, muitos pesquisadores têm investigado a representação da associação de informações por meio de redes ou grafos, chamadas de redes complexas.

Nesse caso, os dados ou as informações, por exemplo, documentos na Web, genes na biologia, ou computadores na malha de servidores são representados em vértices (ou nós), e em arestas (ou ligações), as quais ilustram os relacionamentos entre os vértices. Desde a teoria de grafos apresentada por Euler em 1735, até os estudos de redes de mundo pequeno em 1998, e de redes de livre escala em 1999, cientistas investigam as propriedades e os princípios da criação das redes complexas. A compreensão dessas redes ainda está em fase inicial, e uma grande comunidade de cientistas em variadas áreas está procurando respostas para algumas perguntas como, por exemplo, como pode a estrutura da rede afetar o tráfego na Internet, ou o desempenho de um motor de busca, ou a dinâmica de sistemas sociais? O conhecimento da rede social de uma parte de uma determinada sociedade, por exemplo, pode auxiliar na prevenção da propagação de uma epidemia? Como a indexação associativa humana, vislumbrada por Vannevar Bush, poderia ser representada por uma rede complexa?

Enfim, as informações e suas relações, sejam representadas em redes, ou mesmo como hiperligações, estão presentes em diversos domínios, e são o bem mais valioso da era moderna. Pesquisadores de diversas áreas buscam maneiras de explorar as informações e seus relacionamentos. A Recuperação de Informação (RI) quer ser capaz de unir múltiplas informações ou mesmo modalidades de informações para aprimorar o desempenho da recuperação apoiada na premissa de que a heterogeneidade de múltiplas fontes de informação e/ou algoritmos permite além da complementaridade do conteúdo, a correção de resultados recuperados pelo cruzamento de informações [19]. A Extração da Informação (EI) visa localizar e extrair informações relevantes expressas em linguagem natural de documentos, de maneira automática, e converter as informações extraídas em estruturas que facilitem a manipulação e a análise das mesmas [20, 21]. A Mineração de Texto (MT) é interdisciplinar, e composta por conhecimentos de áreas como RI, EI, Processamento de Linguagem Natural (PLN), Aprendizado de Máquina (AM), Linguística Computacional, entre outras áreas [22]. A MT analisa e busca informações relevantes em textos não-estruturados [23]. A análise de redes de informação está relacionada à MT, à RI (especificamente, a análise de ligações em redes sociais), à Hipermídia e à Mineração da Web e de Grafos.

Esta pesquisadora acredita que uma informação com baixa expressividade de significado é melhor descrita pela sua ocorrência em outras ocasiões em múltiplas fontes de informação ou mesmo em diferentes mídias. Nesse contexto, a pesquisadora Alessandra Alaniz Macedo tem investigado a modelagem e a criação de métodos e mecanismos computacionais para relacionar informações de diferentes tipos advindas de diferentes repositórios de informação. Acessando os resultados providos por esses serviços, usuários podem consultar relatos diferentes de uma mesma informação [24, 25, 26, 27, 28, 3, 10, 29, 6]. Dessa maneira, esses usuários não

precisam formular consultas e ler todo o conteúdo do domínio de informação para encontrar informações relacionadas ao seu assunto de interesse. Serviços similares foram criados na área de Informática Biomédica [11, 2, 12, 30, 9]. Atualmente, a pesquisadora está investigando a representação do seu domínio do problema (informações relacionadas, porém advindas de diferentes fontes em diferentes formatos e mídias), por meio de redes complexas heterogêneas de informação.

Este texto apresenta grande parte do trabalho desenvolvido pela pesquisadora em sua vida acadêmica. O texto foi sistematizado em quatro partes. A primeira parte aborda os problemas de relacionamento de informação textual em repositórios Web, utilizando principalmente RI. A segunda parte extrapola o formato da informação manipulada e o ambiente de captura da informação para definir os relacionamentos entre as informações. A terceira etapa apoia a medicina translacional estendendo os mecanismos de relacionamento de informações para a manipulação de informações biomédicas, explorando principalmente técnicas de MT, PLN e AM. Por fim, a quarta etapa, em execução, está buscando representar os conceitos e os resultados dos trabalhos em grafos semânticos para posteriores análises das propriedades de redes complexas heterogêneas. Além dessa divisão explícita de tema, a mesma divisão segue uma linha temporal, quase exata, referente ao desenvolvimento das pesquisas representando a evolução dos temas. Em todas as partes do texto, pretende-se ilustrar o domínio do assunto abordado, manipulação e relacionamento de informação, bem como as contribuições da pesquisadora nas áreas de conhecimento envolvidas.

Considerando a organização do restante deste documento, o Capítulo 2 aborda a extração, a manipulação e o relacionamento de informação textual, a partir de repositórios Web; o Capítulo 3 apresenta os primeiros trabalhos de relacionamento de informação multimodal e multimídia; o Capítulo 4 descreve o relacionamento e a classificação de informações de pacientes com base em registros e laudos médicos, com o intuito de realizar pesquisa translacional na área da saúde. A apresentação da área de redes complexas e uma discussão da modelagem das investigações sobre relacionamento de informações no contexto de redes complexas são detalhadas no Capítulo 5. Finalmente, a conclusão, no Capítulo 6, resume o caminho científico desta pesquisadora, expondo suas contribuições e alguns aspectos laboriosos de sua carreira acadêmica, os quais demandaram reflexões, aprendizados e mudanças de percursos para trilhar novos desafios.

2 Manipulação e Relacionamento de Informação Textual

As constantes atualizações da informação e o fato da informação ser essencial para o engajamento do cidadão na sociedade e para o desempenho de suas atividades profissionais colocam a informação como foco de atenção de investigações por séculos. Nesse sentido, pesquisadores estudam as necessidades de informação e os usos da informação. Na atualidade, esses estudos buscam principalmente responder à explosão de informações e novas tecnologias que rodeiam as pessoas.

Ao estudar o uso e a necessidade de informação, as seguintes ações são adjacentes à manipulação da informação: selecionar, organizar, interpretar e relacionar. Bush introduziu a noção de associação de conceitos ou de palavras na organização da informação, pois ele acreditava que este fosse o padrão utilizado pelo cérebro humano para transformar informação em conhecimento. Ele indicou que os sistemas de classificação e de indexação precisavam ser intuitivos e não limitadores e que os processos para armazenar e recuperar informação deveriam ser operacionalizados por associação de conceitos como fazem os seres humanos. O Memex, proposto por Bush, armazenava e recuperava documentos por meio de associação de palavras [16].

Buscando associar informações de modo a estender informações iniciais, esta pesquisadora tem investigado o relacionamento ou a associação de informações em diversas mídias, em diversos cenários e para diversos propósitos. Atualmente, tem-se buscado estender essas pesquisas com a análise de propriedades e de topologias de redes complexas nos trabalhos realizados. Em um futuro próximo, a análise da rede poderá incorporar análise de conteúdo de arestas (vértices) e de atributos de elos (conexões). Dessa maneira, pretende-se aprimorar os resultados das associações entre informações.

Na literatura, alguns trabalhos atuais e outros clássicos apresentaram diferentes estilos de manipulação (comparações lexicais, uso de semântica e análise de estrutura) de informação para definição de relacionamentos entre informações textuais [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44].

Neste capítulo, são apresentados os esforços realizados pela pesquisadora em termos de relacionamento de informação textual, por meio da criação de hiperligações na Web. Na Seção 2.1, os trabalhos desenvolvidos durante o doutoramento desta pesquisadora são apresentados. Eles demonstram o início da investigação científica sobre relacionamento de informação. Nas Seções 2.2 e 2.3, são apresentados, respectivamente, os trabalhos de manipulação e relacionamento de informação textual na Web e a busca por similaridade de patentes; ambos realizados após a admissão da pesquisadora na FFCLRP-USP de Ribeirão Preto.

2.1 Criação Automática de Ligações

Durante estágio no exterior, na época de seu mestrado, esta pesquisadora iniciou uma pesquisa de definição e criação de ligações hipertextos entre os documentos Web gerados automaticamente, a partir da captura de experiências ao vivo pelo ambiente educacional ubíquo de captura e acesso, o eClass [45], com os documentos criados a partir do ambiente de edição colaborativa, a CoWeb [46]. Essa abordagem explorava a homogeneidade do conteúdo dos repositórios utilizando comparações lexicais de palavras, em vez de técnicas elaboradas de Recuperação de Informação. O Serviço de Criação Automática de Ligações Lexicais (SCALL) entre informações capturadas pelo eClass e estendidas pela CoWeb foi publicado no *Proceedings do International Workshop on Information Integration on the Web*, em abril de 2001 [47]. Essa pesquisa teve continuidade no doutorado da pesquisadora.

Com a ideia de complementar as ligações lexicais, foi definido o Serviço de Criação Automática de Ligações Semânticas (SCALS) [24]. Esse serviço foi apoiado pela técnica de Indexação da Semântica Latente (*Latent Semantic Indexing* – LSI) [48] que propõe a criação de matrizes de termos por documentos a serem manipuladas pelo modelo algébrico *Singular Value Decomposition* (SVD) [49]. Além de definir ligações entre os ambientes eClass e CoWeb, o SCALS identificou um grande número de ligações entre notícias publicadas por jornais *online* na Web. O Serviço de Criação Automática de Ligações Semânticas e sua infraestrutura foram apresentados na *ACM Symposium on Document Engineering* em Atlanta-USA, em novembro de 2001.

A criação de ligações sobre qualquer repositório Web requer a inclusão de ligações em documentos, portanto, permissões de escrita tornam-se um obstáculo para os serviços automáticos. Em sistemas hipermídia abertos, ligações hipertexto são manipuladas, via interfaces Web independentes, e armazenadas em bases de dados especiais — bases de ligações. A criação de um Serviço Aberto de Criação Automática de Ligações Semânticas (SACALS) foi motivada pelos

resultados obtidos com o serviço SCALS e pela necessidade de armazenar os relacionamentos das ligações identificadas em uma base de ligações que pudesse ser utilizada por outras aplicações, hipermídia ou não [25]. O serviço LinkDigger, criado para instanciar o serviço SACALS, foi apresentado no VIII Simpósio Brasileiro de Sistemas Hipermídia e Multimídia em Fortaleza-Brasil, em novembro de 2002 [50]. O SACALS, uma infraestrutura de suporte e os experimentos definidos foram apresentados na Conferência *ACM Hypertext* em Maryland-USA, em junho de 2002 [25].

Para aprimorar os resultados retornados de acordo com as expectativas dos usuários, conceitos de *relevance feedback*¹ [19] foram explorados na criação automática de ligações semânticas com a participação explícita de usuários [51], após a geração das ligações pelo serviço SCALS. Entre as ligações apresentadas, usuários podem selecionar as que julgarem relevantes para serem consideradas para novos cálculos de similaridade entre os documentos. A pesquisa sobre a participação de usuários na criação de ligações foi apresentada na *XXVIII Conferência Latino-Americana de Informática (CLEI)* em Montevideo-UR, em dezembro de 2002 [51].

Usando os conceitos dos serviços de ligações apresentados nesta seção, o serviço de criação de ligações LinkDigger foi criado e aplicado no contexto de recomendação de páginas Web visitadas por grupo de usuários. Esse sistema visava, assim como Engelbart, alavancar a memória coletiva de grupos de usuários com interesses comuns. A integração LinkDigger aos conceitos de recomendação, denominado WebMemex, foi apresentada na Conferência *ACM Hypertext* em Nottingham-UK, em agosto de 2003 [26]. Em outra oportunidade, esse serviço de recomendação integrado ao LinkDigger foi utilizado como prova de conceito de uma aplicação sensível a contexto suportada por tecnologias da Web Semântica. Esse trabalho foi publicado no *XI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* em 2005 [52]. O WebMedia publica seus artigos completos na *ACM Digital Library* para consulta.

2.1.1 Framework para Criação Automática de Ligações

Os serviços apresentados foram abstraídos em um *framework*² de software chamado de CARE (Criação Automática de Relacionamentos).

O uso dos conceitos de *framework*, na modelagem de componentes do processo de criação automática de ligações em repositórios de conteúdo homogêneo, provê estruturas de projeto e de código a serem reutilizadas por aplicações que tenham o mesmo objetivo. O *framework*

¹ *Relevance feedback* é a reformulação consulta mais popular (expansão ou reponderação) [19].

² Em Engenharia de Software utiliza-se o conceito de *framework* para designar a tecnologia que oferece aos desenvolvedores de aplicações uma maneira poderosa de aumentar a reutilização de componentes de *software*, sendo que inúmeras aplicações distintas podem ser criadas por meio de sua instanciación [53].

CARé é representado por um conjunto de classes, métodos e relacionamentos que dão apoio a tarefas de coleta de informações, pré-processamento textual, geração de termos de índices, relacionamento de documentos, armazenamento, acesso e interações de usuários para o domínio de aplicações com interesse na criação automática de relacionamentos. A criação de um *framework* demandou a definição de estruturas de projeto (classes, métodos, relacionamentos e diagramas) e a implementação de classes e métodos a serem reutilizados por aplicações que tenham o mesmo objetivo.

Para o projeto das estruturas de projeto do CARé, foram definidos um conjunto de termos e seus significados que envolvem a abstração do *framework* em um diagrama de classes apresentado na Figura 2.1. Na apresentação do *framework* é utilizada a seguinte terminologia para a descrição da hierarquia de classes:

- *C* representa a classe controladora “CARé”, que está relacionada com as seguintes classes: *ci*, *pp*, *ir*, *oc*, *ar* e *ac*. A classe *ir* pode ser especializada nas subclasses *rs* e *rl*. “CARé” é na verdade o nome do *framework*;
- *ci* representa a classe “ColetaInformacao”, que possui métodos para realizar a coleta das informações a serem relacionadas;
- *pp* representa a classe “Pré-processamento”, que é formada por um conjunto de métodos que realizam funções de pré-processamento de informações textuais;
- *ir* representa a classe “IdentificacaoRelacionamento”, que identifica os relacionamentos entre informações a partir do espaço semântico definido na subclasse “RelacaoSemantica” (*rs*) ou simplesmente pela comparação lexical dos termos a partir da subclasse “RelacaoLexical” (*rl*);
- *oc* representa a classe “OperacaoConsulta”, que apóia as formulações de consulta e ao método *relevance feedback*;
- *ar* representa a classe “Armazenamento” que manipula chamadas de funções de pacotes que armazenam os relacionamentos identificados;
- *ac* representa a classe “Acesso”, que suporta o acesso às informações similares relacionadas.

Os termos wView e iView são utilizados para representar pacotes externos compostos por aplicações de visualização de informações que podem ser instanciadas a partir da classe “Acesso”. Já os termos XML, MySql e WLS são utilizados para representar pacotes de funções

Figura 2.1: Diagrama de Classes do *framework* CARE para apoio à criação automática de ligações entre repositórios homogêneos de informações [1].

de armazenamento de ligações relacionados à classe “Armazenamento”. Finalmente, o termo *mnoGoSearch* refere-se a um pacote relacionado a classe “Coleta de Informações”.

Cada uma das classes de composição da classe controladora é formada pelos seguintes artefatos de software E_i, r_i, at_i, m_i e S_i obedecendo a seguinte terminologia:

- E_i representa as informações de entrada para uma determinada classe i , sendo que $i = \{I | I \in C\}$;
- r_i representa um conjunto de requisitos da classe i , sendo que $i = \{I | I \in C\}$;
- at_i representa um conjunto de atributos da classe i , sendo que $i = \{I | I \in C\}$;
- m_i o conjunto de métodos que compõem a classe i , sendo que $i = \{I | I \in C\}$;
- S_i representa as informações de saída de uma determinada classe i , sendo que $i = \{I | I \in C\}$. Cabe mencionar que o conjunto de saída de uma classe pode ser o conjunto de entrada de uma outra.

Na definição do *framework* foram utilizadas as infraestruturas definidas na análise do domínio apresentadas como serviços de criação de ligação. Alguns dos processos definidos nessas infraestruturas podem ser diretamente identificados no diagrama de classes. Outros foram agrupados em uma classe que identifica seu comportamento. Por exemplo, o processo “Computa Similaridade” foi diretamente abstraído como a classe “IdentificacaoRelcionamentos”.

O diagrama de classes é composto pelas seguintes classes: classe controladora (CARE – C), classe de coleta de informações (ColetaInformacao – ci), classe de pré-processamento (Pré-processamento – pp), classe de identificação de relacionamento (IdentificacaoRelacionamento – ir), classe de definição de relacionamentos lexicais (RelacaoLexical – rl), classe de definição de relacionamentos semânticos (RelacaoSemantica – rs), classe de armazenamento (Armazenamento – ar), classe de manipulação de consulta e *relevance feedback* (OperacaoConsulta – oc) e classe para acesso às informações similares relacionadas (Acesso – ac). Cada uma dessas classes é composta por métodos, atributos, informações de entrada e de saída.

No diagrama de classes, relacionamentos entre as classes componentes ilustram a associação entre as classes e a comunicação entre os métodos definidos nas classes. Esses relacionamentos são representados por setas contínuas. Por exemplo, na classe “Acesso” (ac) a conectividade com as classes principais do *framework* podem ser interpretadas como, (i) atendimento da consulta gerada pela classe “OperacaoConsulta” (oc), (ii) envio de informação a ser relacionada, e (iii) requisição à classe “Armazena” (ar) para obtenção de relacionamentos armazenados. Já na

classe “IdentificacaoRelacionamento” (*ir*), o relacionamento “identifica”, representa a ativação do processo de ligações da informação coletada ou pré-processada pelos respectivos métodos. As setas tracejadas representam dependência da classe com sistemas externos, como é o caso da classe *ar* quando utiliza a biblioteca de ligações do WLS. O outro tipo de seta representa especializações de classes como, por exemplo, a classe *ir* que é uma generalização das subclasses *rs* e *rl*.

O *framework* CARE foi publicado como tese de doutorado da pesquisadora [54].

2.2 Outros Mecanismos de Criação Automática de Ligações

Após sua contratação como professor-doutor na FFCLRP-USP em agosto de 2004, esta pesquisadora desenvolveu outros trabalhos apresentados no contexto de pesquisas em identificação automática de relacionamentos entre informações.

2.2.1 Criação Automática de Ligações entre Informações Multilíngues

Estudos revelam o uso intenso da Internet em países que não falam inglês [55]. Este cenário cria uma demanda por serviços inovadores de recuperação de informação para melhor apoiar a comunidade mundial. Nesse contexto, o objetivo de *Cross-Language Information Retrieval* (CLIR) é o tratamento de situações quando uma consulta é feita em uma linguagem e a aplicação é capaz de retornar documentos em outro idioma. Muitas técnicas de CLIR exploram a tradução de consultas de usuários, por meio de dicionários de tradução. Entretanto, essas técnicas têm limitações em termos de abrangência lexical do dicionário adotado. Para algumas aplicações, os dicionários são manualmente editados a fim de aprimorar os resultados. Essa atividade requer muito esforço para a representação de uma vasta coleção de informações.

Esta pesquisadora trabalhou na criação de um componente de relacionamento de conteúdo bilingue para a definição automática de relacionamentos entre documentos Web escritos em diferentes idiomas. A abordagem inicial proposta pela pesquisadora estava apoiada na técnica de LSI [49], utilizada nos trabalhos iniciais descritos na Seção 2.1. Para identificar relacionamentos entre documentos em diferentes idiomas, a proposta manipulava a raiz de cada palavra como termo de índice dos correspondentes documentos Web, antes de construir o espaço de informação manipulado pela técnica de LSI. Para experimentar essa investigação, esta pesquisadora e colaboradores estudaram a criação de ligações hipertexto entre documentos advindos de jornais eletrônicos em inglês e em espanhol nas seguintes categorias: lazer, tecnologia e

mundo. Os resultados dessa primeira investigação foram positivos e podem ser encontrado em um artigo publicado no *12th Symposium on String Processing and Information Retrieval* (SPIRE2005) em Buenos Aires-Argentina, em outubro de 2005 [28].

Um caso de estudo apresenta o MedLink, um serviço composto pelo LinkDigger como serviço Web multilíngue integrado ao sistema ArcaMed de captura de informações em reuniões clínicas. O ArcaMed é apresentado no próximo capítulo. O serviço MedLink identifica automaticamente relações semânticas entre casos clínicos multilíngues e torna-os disponíveis para os profissionais da saúde que participam de uma reunião clínica.

Como prova de conceito, casos clínicos multilíngues em Oftalmologia foram relacionados. A coleção de casos de clínicos possui 145 casos clínicos em Oftalmologia coletados de experiências do Brasil (<http://www.scielo.br>) e da Espanha (<http://sciELO.isciii.es>) armazenados na biblioteca eletrônica Scielo (*Scientific Electronic Library Online*). A coleção foi formada por renomados periódicos de Oftalmologia do Brasil e da Espanha, incluindo os *Arquivos Brasileiros de Oftalmologia*, a *Revista Brasileira de Oftalmologia* e o *Archivos de la Sociedad Española de Oftalmologia*. A coleção utilizada foi composta por 69 casos clínicos em Português do Brasil e 76 em espanhol da Espanha.

Cada caso clínico foi coletado da Scielo e formatado como dado de entrada na interface de usuário do ArcaMed para comparação com os casos clínicos em discussão e captura. Apesar de ser manual, esta fase de preparação de captura de reunião clínica requereu também a inserção de informações do caso clínico a ser discutido.

Foram criados 454 relacionamentos entre os casos clínicos e os resultados foram analisados qualitativamente por médicos e docentes com sólida experiência em Oftalmologia e nos idiomas usados para descrever os casos clínicos. Os resultados da análise foram descritos em termos de precisão, revocação e a média harmônica desses dois valores, denominada medida F (ou *F-measure*). A Figura 2.2 possui eixo X que descreve os limites inferiores para filtrar o número de relacionamentos criados e eixo Y que apresenta os valores da medida F, precisão e revocação da execução do MedLink. A figura ilustra que para uma precisão de 100 % obtém-se 0,7 % de revocação — neste caso, apenas quatro documentos são recuperados. O melhor valor de F ocorre quando a filtragem limite é 22 % no eixo X — com 138 associações criadas (19 % de precisão e 26 % recuperação).

Em um mundo globalizado, o MedLink pode ser útil para vários fins, como a investigação, o ensino e a tomada de decisões em cenários de cuidados a saúde, incluindo discussões multiprofissionais para execuções cirúrgicas. A proposta e os resultados foram publicados no *Journal of Universal Computer Science* em 2014 [2].

Figura 2.2: Medida F, precisão e revocação para relacionamentos entre casos clínicos no domínio de Oftamologia a partir da Scielo [2].

2.2.2 Criação Automática de Ligações a partir do Suporte de RSS

Leitores utilizam a Web diariamente para acessar notícias, resultados de máquinas de buscas, novidades, empregos, atividades de lazer, bibliotecas digitais e outros – o rastreamento de tais navegações é uma tarefa bastante complicada para o usuário. O formato RSS (*Rich Site Summary*³) – também conhecido como *RDF Site Summary* ou *Really Simple Syndication* – é um formato XML que permite a publicação de ligações e outras informações, incluindo metadados, que ajuda leitores a decidirem se querem ou não seguir uma ligação disponibilizada. Isso possibilita autores a informarem os seus leitores sobre novos conteúdos.

Esta pesquisadora propôs uma infraestrutura de software para a definição de hiperligações por meio do processamento de documentos Web que incorporam a manipulação de informações RSS. Anteriormente, os serviços de ligações manipulavam matrizes (representando o produto-interno como relacionamentos) que precisavam ser totalmente recalculadas na presença de documentos atualizados, incluindo novos documentos, mudanças e remoções.

Com a proposta de manipulação de RSS, o mecanismo RSS notifica o serviço de ligações sobre as mudanças nos documentos e este reprocessa apenas documentos atualizados. O uso do módulo RSS para criação de ligações foi exemplificado por meio da extensão do sistema de recomendação de informações WebMemex [26], citado anteriormente. Esse trabalho foi publicado no *XI Simpósio Brasileiro de Sistemas Multimídia e Web* (WebMedia 2005) em Poços de Caldas-Minas Gerais, em novembro 2005 [56].

³Informações no W3C incluem <http://www.w3.org/2001/10/glance/doc/howto>. Outras versões estão disponíveis em <http://blogs.law.harvard.edu/tech/rss>

2.2.3 Criação Automática de Ligações a partir de *Clustering* e da Categorização de Informações

Pesquisadores da área de Recuperação da Informação (RI) investigam tecnologias para coletar, representar, organizar, indexar, recuperar, agrupar e classificar grande volume de informação. *Clustering* e classificação (ou categorização) são temas de pesquisa importantes em RI. Usualmente, esses temas são diferenciados da seguinte, uma vez que *clustering* é um método de classificação não-supervisionada. O problema é agrupar objetos, dada coleção de novos elementos não-rotulados, de acordo com rótulos e padrões em agrupamentos significativos. Categorização de informação por meio de sua associação a categorias pré-definidas ou *clustering* buscam a organização de documentos e compõem, no caso de categorias por conteúdo, repositórios com conteúdo homogêneos. Esta pesquisadora acredita que repositórios mais homogêneos de conteúdo de informação possam suportar de maneira mais satisfatória a criação de hiperligações. Com essa motivação, esta pesquisadora realizou duas investigações:

- *Clustering* considerando a Fórmula de Rocchio [57]. Considerando as etapas que o serviço de ligações LinkDigger já realizava, um mecanismo de *clustering* de conteúdo, explorando a fórmula de Rocchio apoiado por coleção de treinamento, foi construído. Esse mecanismo primeiramente pré-processa a coleção de documentos. Assim, todos os documentos passam por um processo de seleção de palavras, remoção de *stopwords* e contagem de frequência de cada palavra em cada documento. Como resultado, obtém-se uma matriz de frequência de palavra por documento. A partir dessa matriz, calcula-se os pesos de cada palavra em cada documento de acordo com esquemas de atribuição de pesos baseado em *tf* (frequência da palavra no documento) e *idf* (frequência da palavra nos outros documentos da coleção). Com a aplicação da fórmula de peso sobre a matriz frequência por documento, o algoritmo proposto gera a matriz de peso por documento e o produto interno corresponde ao cosseno entre eles como uma medida de similaridade. Após a formação da matriz de similaridade entre os documentos, a etapa de formação dos agrupamentos é iniciada. Procura-se o documento mais similar a um determinado documento da coleção, a partir desse momento, esses dois documentos passam a formar um agrupamento e o centróide desse agrupamento é calculado com a fórmula de Rocchio [57]. O novo vetor, centróide, passa a representar os dois documentos no espaço vetorial. Assim, todos os documentos passam por essa etapa e são adicionados a um agrupamento de documentos, ou se não forem similares a nenhum, passam a formar um novo agrupamento. Esse trabalho foi experimentado em uma coleção de artigos e tex-

tos médicos, disponibilizados na Web pela PubMed⁴. A pesquisa, os experimentos e os resultados foram publicados no *XIII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* em 2007 em Gramado-RS, em outubro de 2007 [58].

- Categorização considerando Palavras e o Coeficiente de Jaccard⁵. Para definir categorias com documentos a serem relacionados, um algoritmo proposto por esta pesquisadora suporta a definição de categorias de informação como entrada de dados para a definição de hiperligações lexicais e semânticas. O algoritmo foi implementado e incorporado ao serviço LinkDigger para definir automaticamente categorias de documentos, de acordo com seu conteúdo, a serem utilizadas pelo módulo de criação automática de hiperligações. Para implementação do módulo categorizador foi utilizada a estrutura de referência lexical Wordnet [59]. O Wordnet é formado por objetos lexicais organizados de acordo com a semântica e com distinção entre substantivos, verbos, adjetivos e advérbios. O uso de uma estrutura de conceitos como o Wordnet permitiu a descoberta de diferentes categorias nas quais os termos de índice de cada documento se inserem. Esse trabalho foi publicado no *XII WebMedia* em Natal-RN, em novembro de 2006 [60].

Para dar continuação a esses trabalhos, esta pesquisadora e colaboradores definiram uma infraestrutura de suporte à categorização e *clustering* para definição de hiperligações e realizaram alguns experimentos. Os resultados dos experimentos preliminares com LinkDigger-C são positivos. Em dois experimentos realizados com duas coleções de referência, Cystic Fibrosis e TREC, as ligações criadas eram mais precisas. Esse trabalho foi publicado na Conferência *ACM Hypertext* em Manchester-UK, em setembro de 2007 [61].

2.2.4 Criação Automática de Relacionamentos em Engenharia de Software

A Engenharia de Software é tipicamente uma das áreas da Ciência de Computação que envolve, além de um grande volume de documentos, uma grande diversidade de tipos de documentos tais como diagramas, textos, códigos-fonte, executáveis e etc. A dificuldade de leitura e reúso de informações é parte das preocupações de profissionais de Engenharia de Software com relação à documentação de projeto e de implementação, por exemplo, para a realização das etapas da Engenharia Reversa e Reúso. A possibilidade de encontrar informações relacionadas entre documentações de projeto e de implementação facilita a manutenção e possibilita a

⁴www.ncbi.nlm.nih.gov/pubmed

⁵Se A e B são dois conjuntos de elementos então a similaridade entre A e B é $Sim = |A \cap B| / |A \cup B|$.

reutilização de informações e de experiências. Outro campo de pesquisa da Engenharia de Software que pode se beneficiar da identificação de informações similares é o Controle de Versões.

Motivados por esse contexto, esta pesquisadora procurou aplicar seus trabalhos de relacionamento de informações em documentos advindos da documentação do processo de desenvolvimento de um software médico. Assim, ela utilizou as aulas práticas das disciplinas, ministradas por ela, de Engenharia de Software e de Programação Orientada a Objetos, entre 2004 e 2005, para desenvolver e documentar as fases de desenvolvimento de um sistema de informação para apoiar as atividades do Centro Médico Social e Comunitário de Vila Lobato (CMSCVL) em Ribeirão Preto/SP. Foram aplicados conceitos de especificação, projeto, codificação e testes de software apoiados por tecnologias como a meta-linguagem UML, ambientes de apoio ao desenvolvimento de projeto e a linguagem de programação Java. Esse projeto envolveu estudantes de informática biomédica e a participação de outros professores do curso e do CMSCVL. O trabalho desenvolvido para o CMSCVL foi redigido como poster e foi apresentado no 13º Simpósio Internacional de Iniciação Científica da USP (SIICUSP), em novembro de 2005 [62]. Outra parceria com o CMSCVL é apresentada no Capítulo 4.

Considerando toda a documentação gerada nesses dois anos do projeto CMSCVL, esta pesquisadora e seus orientados estenderam o serviço de criação de ligações para definir relacionamentos entre as informações de documentação, de especificação de requisitos, de projeto, de codificação e de testes do sistema de informação. Esse projeto está dividido em duas etapas:

- Identificação Automática de Relacionamentos entre Informações Similares advindas da Documentação de Projetos de Software. Em um projeto de pesquisa PIBIC (109796/2005-5) de agosto de 2005 a janeiro de 2007, foi realizada a adaptação e a integração de alguns módulos do serviço de criação automática de relacionamentos com o intuito de aplicá-lo no contexto de documentação de projeto. Para alcançar esse objetivo, os módulos foram adaptados e integrados a uma ferramenta de apoio à documentação de software denominada MVCASE⁶. Assim as experiências de projetos anteriores suportados pela ferramenta podem ser reutilizadas por outros desenvolvedores. Usuários da ferramenta MVCASE podem, durante o desenvolvimento de seus sistemas, encontrar e reutilizar experiências de outros desenvolvedores de software em seus projetos. Após inicializado o processo de relacionamento de informações, o usuário é contemplado com uma lista de documentos relacionados ao documento em edição. Nessa lista, o usuário pode escolher o documento relacionado de interesse e visualizá-lo na parte inferior da mesma interface. Além de reúso, a possibilidade de encontrar informações relacionadas entre

---

⁶<http://mvcase.dev.java.net>

documentações de projeto e de implementação pode facilitar a manutenção, que é responsável por praticamente 70% do custo de desenvolvimento de software. A pesquisa, o desenvolvimento e os resultados deste trabalho foram aceitos para publicação no *XII Simpósio Brasileiro de Sistemas Multimídia e Web* (WebMedia em 2006) em Natal-RN, em novembro de 2006 [63].

- Identificação Automática de Relacionamentos entre Conteúdo Extraído de Imagens. Além de envolver uma grande quantidade de documentos, a área de Engenharia de Software manipula uma grande diversidade de documentos compostos por figuras tais como diagramas e códigos-fonte. Portanto torna-se muito importante o tratamento e a recuperação de, por exemplo, imagens e códigos de programa para reúso de informações apresentadas em documentações de projeto. O objetivo da segunda fase desse projeto era dar continuidade ao desenvolvimento da nova versão da ferramenta MVCASE por meio da criação (i) módulos de tratamento diferenciado e recuperação de imagens e códigos de programa e (ii) novas interfaces de usuários. Esse trabalho também foi desenvolvido no contexto de um projeto PIBIC de *fevereiro a julho de 2007*.

Ainda em termos de divulgação científica a partir de trabalhos realizados no contexto de sala de aula em 2005, esta pesquisadora orientou o desenvolvimento de uma investigação na disciplina de Estrutura de Dados que gerou a publicação de um artigo de iniciação científica no WebMedia 2005, realizado em Poços de Caldas-MG [64]. O objetivo foi um estudo comparativo, a partir da aplicação de diferentes algoritmos e estruturas de dados, dada a manipulação de grandes volumes de informações textuais no contexto biomédico. Como resultado do estudo foi gerada uma tabela comparativa considerando a complexidade, o desempenho e a eficiência de cada estrutura de dados e seus respectivos algoritmos experimentados.

2.3 Busca Guiada de Patentes

Pesquisas recentes demonstram que as patentes armazenam grande parte do conhecimento mundial [65]. As patentes disponibilizam informações técnicas mais detalhadas que artigos científicos, auxiliam na divulgação do conhecimento e estimulam o desenvolvimento tecnológico de diversas áreas [66]. No entanto, esses documentos são pouco utilizados como fontes de referência no meio acadêmico, por serem considerados pouco acessíveis. As ferramentas públicas de busca de patentes são limitadas, a linguagem utilizada nas redações é complexa e para patentes de áreas multidisciplinares, como a bioinformática, a complexidade é maior para usuários de domínio disciplinar. Estudos recentes demonstram que a utilização de mecanismos como

classificadores e artefatos linguísticos auxiliam na obtenção de resultados mais relevantes aos usuários de informações de patentes [65].

Esta pesquisadora e um aluno de mestrado exploraram artefatos linguísticos, técnicas de Recuperação de Informação e de Classificação Textual para criar o Sistema de Busca Guiada de Patentes de Bioinformática (BPS). O BPS é um classificador automático de informações de patentes em bioinformática que filtra as informações retornadas. Para modelar o classificador do BPS, foi definido o seguinte processo de descoberta de conhecimento do tipo KDD (*Knowledge Discovery in Databases*):

1. Seleção: obtenção da visão lógica de cada documento pela identificação da estrutura dos documentos da coleção e a definição dos campos a serem coletados;
2. Pré-processamento: eliminação de caracteres especiais e *stopwords* e a redução de cada termo ao seu radical;
3. Transformação: padronização e normalização dos valores de cada atributo para adaptá-los aos classificadores utilizados na etapa de mineração;
4. Mineração de Texto: teste dos algoritmos de classificação multirrótulo (*Random k-Labelsets* (RakEL) e *Multilabel k-Nearest Neighbors* (MLkNN));
5. Avaliação: utilização da Validação Cruzada Estratificada de 10 partes (*10-Fold Cross-Validation Stratified*) como método de validação.

O filtro do BPS processa as consultas dos usuários, utilizando a ontologia *Gene Ontology*⁷ [67] como artefato linguístico e calcula os graus de similaridade entre a consulta do usuário e as classes de Bioinformática. A saída do filtro é uma lista ordenada decrescente pelo grau de similaridade da consulta e das classes, ou seja, as classes mais próximas do topo da ordenação estão mais relacionadas à consulta do usuário. Em seguida, o filtro calcula a similaridade da consulta original com cada documento das duas primeiras classes da lista ordenada de retorno do BPS para a consulta. O filtro do BPS apresenta esse resultado como relacionamentos entre consulta e documentos das classes mais similares.

O sistema BPS é uma ferramenta gratuita, que possui como diferenciais o foco na área de Bioinformática, e a busca guiada por informações de classificação. Essas características tornam o BPS uma alternativa para pequenas e médias empresas brasileiras de pesquisa que precisam fazer buscas por patentes de Bioinformática e não podem arcar com os altos custos de uma

---

⁷<http://www.geneontology.org/>

ferramenta comercial de busca de patentes. Da maneira como foi projetado, o BPS permite ser aplicado a outras áreas do conhecimento. Para isso, o classificador deve ser treinado com uma coleção da área que se deseja aplicar o BPS e incorporar a utilização de outro artefato linguístico (ontologia, tesauro, dicionário, etc) também relativo à essa nova área de aplicação. O sistema BPS utiliza a informação de classificação para melhorar as buscas específicas. O BPS pode ser utilizado como complemento em uma das etapas de investigação de anterioridade no processo de pedido de patentes. O BPS também pode identificar em qual classe sua patente se encaixa melhor e encontrar patentes relacionadas à sua na área de Bioinformática. Essa característica também permite que usuários com patentes de outras áreas possam encontrar patentes de Bioinformática relacionadas a sua.

A coleção de patentes de Bioinformática utilizada como exemplo de classificação para experimentação do BPS foi obtida no site do USPTO ⁸ (*United States Patent and Trademark Office*). Para coletar os documentos foi desenvolvido um *Web crawler*, que acessa a página de buscas avançadas, processa as páginas HTML e coleta automaticamente as informações textuais de cada patente das cinco classificações americanas reservadas para Bioinformática. Para realizar os experimentos, optou-se por trabalhar com uma coleção fixa de documentos coletados até outubro de 2012.

Para avaliar as buscas do BPS, efetuou-se uma análise comparativa entre os resultados deste sistema e os das ferramentas de busca *PatFT* (USPTO), *Patentscope* (WIPO), *Derwent Innovations Index* (Thomson Reuters) e *Google Patents*, as quais indexam a coleção americana de patentes. Para uma análise mais equitativa, critérios de busca iguais para todas as ferramentas foram estabelecidos, com limitação de 256 caracteres para o texto de composição da consulta e delimitação do campo de busca apenas às classes de Bioinformática. Durante a preparação do experimento, observou-se que as ferramentas *Derwent* e *Patentscope* não permitiam delimitar as buscas às classes de Bioinformática. Assim, apenas as ferramentas *PatFT* (USPTO) e *Google Patents* foram comparadas ao BPS em termos de precisão e as buscas no *Derwent* e no *Patentscope* foram delimitadas às classes de Bioinformática da taxonomia internacional IPC1 e exibidas para demonstração do processo de busca nessas ferramentas.

Os algoritmos de classificação multirrótulo *RAkEL* e *ML-kNN* foram experimentados e avaliados com diferentes parametrizações. O algoritmo com parametrizações de melhor desempenho foi o *RAkEL* com *Naive Bayes Multinomial*, o qual foi incorporado como padrão do sistema BPS.

Em relação ao esperado aprimoramento da classificação com o uso da *Gene Ontology*

---

⁸<http://www.uspto.gov>

(GO), os experimentos realizados não conseguiram identificar essa contribuição em termos semânticos. Essa característica pode ser em decorrência dos termos GO serem muito específicos e apresentarem muitos termos não significativos para título e resumo de patentes, ou seja, os termos indexados pela GO. Afinal, esses termos são pouco utilizados na redação do título e resumo de patentes. No processamento de similaridades entre a consulta e os documentos da coleção, observou-se que a Gene Ontology contribuiu na ordenação dos resultados. Em três de quatro casos, as patentes retornadas de classes diferentes da classe da patente buscada possuem maior distância dos primeiros lugares na ordenação. O sistema BPS pode ser estendido para outros domínios de busca de patentes, desde que os artefatos linguísticos usados sejam substituídos pelos artefatos do domínio em questão.

O classificador de documentos de patentes, que atua como filtro no sistema de busca guiada de patentes de bioinformática foi publicado e apresentado em 2013 no Workshop de Informática Médica no Congresso da Sociedade Brasileira de Computação [68]. A continuidade desse trabalho deverá suportar publicações em periódico internacional.

2.4 Considerações Finais

Este capítulo foi iniciado com a apresentação dos trabalhos desenvolvidos principalmente pela pesquisadora, usando RI (comparações lexicais, LSI, CLIR, recomendação) e a Web (RSS e *linkbases*), na fase de seu doutoramento. Os resultados foram abstraídos para a geração do *framework* de software CARE. Nessa fase, o foco da pesquisadora e sua orientadora de doutoramento era a publicação de pelo menos um artigo internacional por ano de doutorado. O objetivo foi atingido, mas extrapolando os objetivos, artigos foram também publicados em eventos nacionais da área de Hipermídia e Web.

Os três últimos trabalhos deste capítulo já possuem a participação de alunos, orientados pela docente. O antepenúltimo e o penúltimo trabalhos envolveram alunos em trabalhos de sala de aula e de iniciação científica, com bolsa PIBIC. Os trabalhos foram publicados no contexto do SIICUSP e de encontros de iniciação científica no WebMedia. Esta pesquisadora considera o projeto com o Centro Médico Social e Comunitário de Vila Lobato uma iniciativa fantástica que levou alunos a executarem seus projetos de sala de aula na prática profissional do informata biomédico em um dos seus campos de atuação. Esse projeto iniciado com alunos de graduação em 2004 gerou excelente resultados, inclusive com projeto de mestrado e publicações internacionais em eventos e em periódicos até 2012. O último trabalho apresentado foi realizado no contexto de um trabalho de mestrado que está sendo continuado por esta pesquisadora.

3 Manipulação e Relacionamento de Informação Multimídia e Multimodal

Informação multimídia é a informação digital representada por diferentes mídias interligadas. Muitas pessoas relacionam informações multimídia com entretenimento, porém informação multimídia está presente em diferentes situações e domínios. Por exemplo, pesquisas científicas no domínio de saúde e de medicina consomem e geram grandes quantidades de dados multimídia. Esses dados possuem diferentes mídias, desde informações textuais simples, imagens com/sem volumes 3D, como tomografias computadorizadas e ressonâncias magnéticas. A partir de hipóteses, os pesquisadores e profissionais da saúde executam experimentos usualmente com os dados obtidos de exames.

Alguns pesquisadores diferenciam informação multimídia de informação multimodal, uma vez que os sistemas multimídia lidam com a apresentação de informações e os sistemas multimodais interpretam e recriam informações apresentadas em diferentes mídias [69]. Alguns pesquisadores apoiam a diferenciação na interface do usuário. Para os pesquisadores, a distinção entre interfaces multimídia e interfaces multimodais é a entrada e a saída do sistema [70]. Assim, uma interface multimodal suporta múltiplas entradas e múltiplas saídas simultâneas de dados, por exemplo, o uso de discurso juntamente com gestos e anotações com tinta digital em lousas eletrônicas. A interface de sistemas multimídia suporta várias saídas, por exemplo, um texto com a informação de áudio ou tátil. Portanto, sistemas multimídia podem ser considerados um subconjunto de pesquisa multimodal [70]. Na literatura, existem outras diferenciações para sistemas multimídias e multimodais [71].

Para gerar conhecimento a partir dos diferentes tipos de dados, a manipulação de informação multimídia ou multimodal torna-se essencial. Neste capítulo, a manipulação e a recuperação de informação multimídia ou multimodal é apresentada com foco na extração de informação semântica dos dados. No caso de informação multimídia, esta pesquisadora tem manipulado principalmente informações textuais com imagens médicas no contexto de tomografia de pulmão e imagens citológicas de sistemas CAD e PACs, imagens geradas a partir de texto, nos

casos de manipulação do ArcaMed, e de imagens de OCR (*Optical Character Recognition*). Para manipular essas mídias, foram investigados e estendidos métodos de extração de característica em imagens, filtragem, classificação e outros. Em informações multimodais, esta pesquisadora definiu o relacionamento de informações textuais com informações extraídas de tinta digital e da fala de usuários, em ambientes ubíquos instrumentados para captura de atividades em sala de aula. Em relação a vídeo, esta pesquisadora contribuiu com um trabalho de relacionamento de texto em vídeo com um grupo de pesquisadores do ICMC (Instituto de Ciências Matemáticas e de Computação) da USP de São Carlos e desenvolveu mecanismos para capturar vídeos em casas inteligentes. As pesquisas e os resultados de definição de relacionamento entre informações multimídias ou multimodais são apresentados neste capítulo.

3.1 Relacionamento de Informação Multimodal

As pesquisas em computação ubíqua estudam como integrar, de forma contínua e não-intrusiva, a tecnologia computacional às atividades do cotidiano de indivíduos [72]. Sob esse paradigma computacional, a interação entre usuários e computadores se aproxima da maneira como as pessoas se comunicam umas com as outras (escrita, fala, gestos). Além disso, as ações implícitas e explícitas que ocorrem nessa comunicação são utilizadas como dados de entrada para sistemas de computação ubíqua.

A computação ubíqua visa proporcionar serviços aos usuários em diferentes contextos e ambientes do cotidiano como o lar, uma sala de reunião ou uma sala de aula. Um tema de pesquisa nessa área é o da construção de aplicações de captura e acesso, as quais suportam a captura (gravação) de informação multimodal gerada durante uma experiência ao vivo e criam documentos multimídia para reproduzir o conteúdo capturado. A captura de informação gerada ao vivo demanda ambientes instrumentados com dispositivos como microfones, câmeras, sensores e lousas eletrônicas. Cada experiência capturada nesses ambientes pode estar relacionada a outras (por exemplo, várias reuniões de um projeto), gerando oportunidades de reúso. Portanto, existe uma demanda por mecanismos para definir automaticamente relacionamentos entre os documentos relativos a diferentes experiências. Esta pesquisadora tem desenvolvido trabalhos de relacionamento de informação multimodal em ambientes ubíquos de sala de aula, reuniões clínicas e casas inteligentes, apresentados a seguir.

3.1.1 Contexto Educacional

A computação ubíqua aplicada à educação pode proporcionar uma aproximação da tarefa cotidiana de aprendizado a tecnologia, além de permitir personalização de ensino. A aproximação de tecnologias tem motivado crianças e jovens a participarem mais ativamente das tarefas que lhes são propostas.

Esta pesquisadora e seu grupo de colegas pesquisadores do laboratório Intermedia do ICMC-USP criou a CALiSP (*Infrastructure for Capturing, Accessing, Linking, Storing and Presenting information*) que compreende: (i) a automação do processo envolvido na fase de captura de aplicações típicas de captura e acesso; (ii) o processamento e armazenamento de informações capturadas como documentos XML; (iii) a identificação de relacionamentos semânticos entre o conteúdo capturado; (iv) o armazenamento das associações identificadas em uma *link-base* aberta; e (v) a definição das folhas de estilo, que permitem aos usuários a visualização do conteúdo capturado em uma variedade de formatos de apresentação [3]. Esta integração para visualização gerou uma nova abordagem para modelagem de informações contextuais. A CALiSP fornece suporte para captura e acesso de informações antes, durante e depois de experiências ao vivo, e abre oportunidades para o uso a longo prazo de aplicações de captura e acesso.

A infraestrutura da CALiSP da Figura 3.1 possui basicamente um *Document Builder* que processa informação capturada como documento XML, o *Link Builder* que identifica relacionamentos semânticos entre informação capturada, o *Storage Manager* que armazena ambos, os documentos correspondentes às informações capturadas e os relacionamentos identificados, e um *Presentation Manager* que explora folhas de estilo para gerar documentos de apresentação. Depois de algum tempo de uso da CALiSP, esta pesquisadora apresentou operações na Web para serem usadas para identificar as ligações. As operações devem ser disponibilizados *antes*, *durante* e *depois* de uma apresentação ao vivo e, em qualquer caso, o usuário deve anexar resultados relevantes como anotações. A ativação das operações pode ser automaticamente realizada por interações multimodais, digitação e navegação. A CALiSP foi apresentada no periódico internacional Multimedia Tools and Applications (MTAP) em 2008.

Motivada pela demanda por interações naturais em aplicações de computação ubíqua e a aplicação da descoberta de relações intrínsecas entre informações em muitas áreas, esta pesquisadora e colaboradores propuseram as seguintes interações multimodais: *linking by typing*, *linking by inserting* e *linking by capturing* [29]. *Linking by typing* ocorre quando um usuário fornece metadados ou palavras-chave à aplicação que usando esses dados relaciona os metadados com outras informações. Normalmente, esse tipo de *linking* ocorre na fase de preparação da

Figura 3.1: Infraestrutura da CALiSP (*Infrastructure for Capturing, Accessing, Linking, Storing and Presenting information*) [3].

sessão de captura. O *linking by inserting* ocorre pela inclusão de slides ou documentos preparados, normalmente na preparação da sessão de captura. Finalmente o *linking by capturing* ocorre desde a captura de informações textuais de slides, de escrita sobre slides ou mesmo do áudio durante as apresentações ao vivo. Durante a fase de acesso, usualmente ocorre *linking by typing*. O modelo subjacente a proposta das interações multimodais faz uso das operações de busca e de recomendação para proporcionar alternativas de ação para relacionar sessões capturadas em ambientes ubíquos. A proposta de diferentes tipos de *linking* para interações multimodais foi apresentada e publicada no *WebMedia* em 2010.

Vale ressaltar que todas as investigações de criação de relacionamentos em ambiente ubíquo no contexto educacional foram desenvolvidas em colaboração com pesquisadores do Laboratório Intermedia do ICMC-USP de São Carlos. Essa parceria de trabalho no domínio educacional propiciou a participação da pesquisadora como colaboradora na Fase I do Projeto TIDIA-Ae e como coordenadora de pesquisa do seu grupo em seu laboratório na Fase II do Projeto TIDIA-Ae, ambos financiados pela FAPESP¹.

¹ A pesquisadora recebeu o valor de R\$ 203.239,06 na Fase II do Projeto TIDIA-Ae, sem considerar os valores pagos a título de bolsa de treinamento técnico aos bolsistas do projeto.

3.1.2 Contexto de Saúde e de Medicina

A área médica tem também se beneficiado com a utilização de sistemas de computação ubíqua. Quando, por exemplo, uma situação de emergência é detectada em um paciente com problemas cardíacos crônicos, um sistema de monitoramento remoto de sinais vitais de pacientes notifica equipes de serviço mais adequadas para a situação, como médicos de plantão e serviço de ambulância. Esta pesquisadora tem participado da definição e da construção do *framework* de software ArcaMed para apoiar o desenvolvimento de aplicações para medicina com tecnologias de computação ubíqua. Uma outra oportunidade é o uso de computação ubíqua para registrar atividades de equipes médicas durante um processo cirúrgico para fins de ensino-aprendizagem de residentes em hospitais.

No contexto de apoio ao diagnóstico médico, o *framework* de software ArcaMed apoia o desenvolvimento de aplicações para medicina com tecnologias de computação ubíqua. O ArcaMed provê o uso de elementos textuais e multimídia que tornam possível a interação de usuários no sentido de manipular e estender registros clínicos com informações de sinais vitais, análises laboratoriais e sintomas clínicos. Essa extensão é possível via dois mecanismos principais: anotações multimídia sobre imagens e vídeos de exames clínicos e associações hipermídia entre registros clínicos e outros documentos relevantes. O projeto inicial do *framework* ArcaMed e a prova de conceito no contexto de apoio ao diagnóstico em radiologia foram publicados no *Workshop de Informática Médica (WIM)* [73] e no *Congresso de Tecnologia e Humanização na Comunicação em Saúde (ConTIC-Saúde)* em julho de 2007 [74].

Ainda no domínio de saúde e de medicina, especificamente de reuniões clínicas, o ArcaMed suporta a documentação multimodal desse tipo de reunião, que normalmente engloba a discussão de casos clínicos graves e de difícil tomada de decisão. Essas reuniões são previamente agendadas e os casos clínicos da reunião são geralmente escolhidos de acordo com a sua gravidade em uma fila constantemente atualizada. Em tais experiências, uma equipe multidisciplinar, que inclui médicos de diferentes especializações, enfermeiros, residentes e alunos de pós-graduação discutem todos os dados pré e pós-cirúrgicos. Nesse contexto, o ArcaMed captura e sincroniza o conjunto de material apresentado e as interações dos profissionais correspondentes, e automaticamente relaciona casos clínicos de pacientes para recomendação e gera documentos acessíveis via Web com todas as informações capturadas [75]. A documentação resultante pode ser usada para vários fins, como ensino, pesquisa e tomada de decisão pré-cirúrgica.

Especificamente no contexto de reunião clínica para suprir a falta de uma documentação adequada dessas reuniões, foi definida a arquitetura de software da Figura 3.2 para a pri-

meira versão do ArcaMed. Esta arquitetura considera cada fase de um sistema de captura e acesso: preparação, captura, pós-produção e acesso. Durante a pré-produção de uma reunião clínica, para cada novo caso clínico (ver Figura 3.2.1), um documento XML é criado com a informação correspondente. Esse documento é enviado para o módulo de associação para encontrar semelhanças entre casos clínicos (ver Figura 3.2.2). Por razões de independência de rede, todos os documentos XML de casos clínicos são armazenados na base de dados do protótipo (ver Figura 3.2.3). Tanto a informação de casos clínicos de pacientes, seus exames e relatórios de investigação são registrados em um documento XML para a reunião clínica (ver Figura 3.2.4), bem como na base de dados do protótipo. Quando um conjunto de casos clínicos é escolhido para fazer parte de uma reunião clínica, estes casos clínicos são enviados para o módulo de associação (ver Figura 3.2.5) para que ele retorne uma lista de casos clínicos relacionado com cada caso utilizado como entrada (ver Figura 3.2.6). Esse trabalho de colaboração inicial em torno da arquitetura do ArcaMed e do primeiro protótipo geraram resultados publicados e apresentados no *XIII Simpósio Brasileiro de Sistemas Multimídia e Web* (WebMidia 2007) em outubro de 2007 [4] e no *Proceedings of the eighth ACM Symposium on Document Engineering* (DocEng) em 2008 [75].

Figura 3.2: Arquitetura ArcaMed [4].

Para impulsionar a concepção e o desenvolvimento de uma nova versão do ArcaMed, foi utilizada etnografia² para levantamento de requisitos em reuniões clínicas [77]. A etnografia é

²A etnografia é um método de pesquisa para compreender como as pessoas realizam suas tarefas em ex-

um poderoso meio de entender o conhecimento tácito sobre necessidades, comportamentos e rituais. Foram acompanhados mais de 20 dias de trabalho de observações de reuniões clínicas de diferentes especializações em Medicina considerando tarefas realizadas antes, durante e depois de tais experiências [77]. Na nova versão, um dos trabalhos deu enfoque na documentação automática dos usuários interações com imagens médicas digitais em formato DICOM [78].

A segunda versão do ArcaMed destacou o uso de mecanismos para a extensão da informação capturada de cuidados de saúde em ambientes ubíquos. Para alcançar tal objetivo, o ArcaMed foi integrado ao serviço de relacionamento de informações semânticas, que identifica e cria automaticamente relações semânticas entre informações de pacientes [79]. Os relacionamentos criados associam informações capturadas antes, durante e após as atividades de reuniões clínicas. Esse tipo de iniciativa foi considerada útil para esses profissionais, que comumente tentam relacionar mentalmente casos clínicos para alcançar um tratamento mais seguro e eficiente de cuidados de saúde a um paciente em análise na reunião. O ArcaMed a partir de etnografia foi publicado no *ACM Symposium on Applied Computing* (ACM SAC) [77] em 2008, com enfoque em DICOM no Workshop de Informática Médica [78] em 2008, e a segunda versão do sistema no *IEEE International Conference on Computational Science and Engineering* (IEEE SEC) [79] em 2008.

3.1.3 Contexto de Casas Inteligentes

Os sistemas de controle domésticos (*Home Control System – HCS*) estão se tornando mais comuns e parte integrante de habitações modernas. O controle computadorizado de alarmes, de sistemas de climatização e de outras aplicações para habitações são tecnologias que podem favorecer residências em todas as classes sociais. Algumas tecnologias, como os sistemas para controle de iluminação, estão presentes em casas, apartamentos e escritórios de médio e alto padrão, além de grandes empresas, teatros, hotéis e hospitais.

Apesar de HCS não estar avançado nos conceitos da Computação Ubíqua, a ideia de ambientes inteligentes se direcionada para esse sentido. Os HCSs ainda são compostos, em sua grande maioria, por sensores concentrados em um computador central de controle, descaracterizando a conectividade de seus componentes. Porém, as interfaces de usuário e de *hardware* se tornam cada vez mais invisíveis em HCSs; não sendo necessário ao usuário conhecimentos em computação para acionar seus comandos. Especificamente, as interfaces de usuários são desenvolvidas para Internet ou mesmo especificamente para Web. Nesse contexto, recentemente esta pesquisadora ministrou um curso sobre padrões Web utilizados para o desenvolvimento de casa experiências do mundo real [76]

inteligentes. Esse curso deu origem ao capítulo de livro publicado no WebMedia 2013 [80]. Este capítulo apresenta e exemplifica o uso dos padrões Web HTML5, CSS3, JQuery e JQueryMobile para o desenvolvimento de interfaces de usuário. Essas tecnologias auxiliam a visualização e o controle de informações estáticas e dinâmicas, por exemplo, em casa inteligente. Como estudo de caso, o capítulo discute e aplica tecnologias na construção de um aplicativo para controle de automação de residências.

A área da saúde mostra-se necessitada da automação de processos, visto que a maioria dos locais de prestação de serviços e administrativos ainda realiza procedimentos de forma manual, dificultando o controle de dados e o gerenciamento de informações. O fato de conhecer as tecnologias de casas inteligentes e de trabalhar com computação ubíqua nos contextos de educação e de medicina, gerou um contato da pesquisadora com um grupo de docentes do Departamento de Neurociências e Ciências do Comportamento da FMRP, especificamente do curso de Terapia Ocupacional (TO). Em reuniões iniciais, a pesquisadora e as docentes estão planejando a escrita de um projeto sobre o desenvolvimento de ambientes inteligentes para idosos ou mesmo para pessoas com problemas mentais do Hospital Santa Teresa, onde as docentes prestam serviços. Uma colaboração inicial foi estabelecida via contribuição desta pesquisadora na organização e na realização do I Congresso de Gereontecnologia³ proposto pelas docentes do curso de Terapia Ocupacional da FMRP-USP de Ribeirão Preto. Esta docente participou da comissão científica e organizou a sessão “Ambientes assistidos para adultos mais velhos” do evento em abril de 2016.

3.2 Relacionamento de Informações de *Optical Character Recognition* (OCR)

Devido à utilização generalizada e multiúso de imagens de documentos e a disponibilidade atual de um elevado número de documentos em formato de imagens, sistemas de recuperação de informação têm sido cada vez mais exigido. Por exemplo, a Biblioteca Clendening publicou imagem digital de textos de história nacional para estimular o uso educacional dessas imagens [81]. Milhões de patentes e pedidos de patente são indexados a partir de imagens de textos nos Estados Unidos para que usuários da Internet possam pesquisar e percorrer as páginas, e fazer manipulação de zoom em áreas de imagem [82].

A fim de apoiar cruzamento de informações históricas, o governo brasileiro digitalizou milhares de registros criminais produzidos durante a ditadura brasileira [83]. Um consórcio

---

³<http://gerontec.wix.com/congresso>

Figura 3.3: Visão geral do LinkDI para geração de relacionamentos considerando imagens de documentos [5] [6].

de bibliotecas da Universidade de São Paulo possui literatura rara a qual foi digitalizada para distribuir seu conteúdo [84]. O uso de bases de dados de imagens de documentos também exige formas robustas de indexação de conteúdo e recuperação.

Esta pesquisadora e colaboradores desenvolveram um método para extrair conteúdo textual de imagens e gerar automaticamente relacionamentos de informações entre documentos do tipo imagem gerado a partir de OCR (*Optical Character Recognition*) [5]. O LinkDI extrai o conteúdo de imagens de documentos índices, e calcula sua semântica latente e as relações entre imagens como hiperligações, conforme a Figura 3.3.

O LinkDI foi experimentado e seu desempenho foi avaliado com base na comparação da qualidade dos relacionamentos criados entre documentos textuais, bem como entre as suas respectivas imagens de documentos. Considerando a coleção de imagens de documentos, foram realizados experimentos para comparar o desempenho de LinkDI quando se explora ou não a técnica de LSI (*Latent Semantic Indexing*). O objetivo mais desafiador de OCR é resolver o falso reconhecimento de caracteres [85]. Os resultados experimentais mostraram que a LSI

pode mitigar o efeito do não-reconhecimento ou do falso reconhecimento de OCR. Esse resultado reforçou a viabilidade de LinkDI para OCR com alta degradação. O LinkDI, sua modelagem e resultados foram publicado no *ACM Symposium on Applied Computing* (ACM SAC) em 2010 [5] e no periódico internacional *Journal of Universal Computer Science* (JUCS) em 2011 [6].

3.3 Relacionamento de Informações Textuais e de Imagens

Alguns motores de busca realizam a busca de imagens, utilizando os textos ao redor da imagem, suas legendas ou nomes de arquivos como seus termos de índices. No entanto, essa prática limita a capacidade dos motores de busca recuperarem as imagens semanticamente relacionadas a uma determinada consulta. O atual “estado da arte” na recuperação de imagens baseada em conteúdo está progredindo, mas ainda não conseguiu fazer a ponte entre a semântica dos conceitos humanos, por exemplo, as consultas baseadas em palavras-chave, e o baixo nível das características visuais extraídas das imagens. Nesse contexto, tornou-se uma necessidade urgente o desenvolvimento de novos paradigmas para a recuperação de imagens, a partir de informação textual. A área de saúde tem esse mesmo apelo em suas aplicações que manipulam imagens: PACS (*Picture Archiving and Communication System* - Sistema de Comunicação e Arquivamento de Imagens) e SADM (Sistema de Apoio ao Diagnóstico Médico). A pesquisadora tem investigado o cenário de recuperação de imagens, a partir de informação textual para esses dois tipos de aplicações. Suas contribuições são apresentadas nas próximas subseções.

3.3.1 Contexto de *Computer-Aided Diagnosis* (CAD)

Sistemas de Apoio ao Diagnóstico Médico (SADM) (*Computer-Aided Diagnosis* - CAD) podem ser usados para ajudar os radiologistas na tarefa de analisar uma imagem médica e estabelecer o diagnóstico final, oferecendo-lhes uma segunda opinião e apontando estruturas de interesse. Além de prover suporte ao processamento de imagens, SADMs podem apoiar a recuperação de informação, utilizando os cabeçalhos em formato DICOM (*Digital Imaging and Communications in Medicine*) e dados inseridos pelo radiologista. Em conjunto com o processamento de imagens, técnicas de recuperação de informações podem ser úteis para eliminar hipóteses e fornecer informações relevantes durante o diagnóstico, agregando conhecimento ao SADM. SADMs podem ser usados na tentativa de aumentar a acurácia do laudo e a consistência de interpretação da imagem, sendo o diagnóstico final sempre feito pelo radiologista.

Existem cerca de 180 patologias classificadas como Doenças Pulmonares Difusas (DPD),

das quais grande parte apresenta características muito similares e geralmente são acompanhadas de sintomas não específicos [86]. A modalidade de exame preferencial para o diagnóstico desse grupo de patologias é a Tomografia Computadorizada de Alta Resolução (TCAR), pois em muitos casos o uso de raio-X convencional não revela a presença da lesão [87]. Devido ao grande número de imagens obtidas por TCAR, o diagnóstico torna-se um processo laborioso e sujeito a erros de interpretação. A variação de combinações de padrões radiológicos complexos e a falta de correlação com os achados patológicos podem levar ao erro [88].

O diagnóstico de DPD é fortemente relacionado com as propriedades de textura do tecido afetado, devido ao caráter difuso destas patologias [86]. Pesquisadores indicam que o uso de atributos de textura e espectrais tem se mostrado eficiente na detecção de DPD [87, 88]. Esta pesquisadora e colaboradores desenvolveram um *framework* de classes de software que suporta a definição e a criação de um SADM para a análise quantitativa e classificação automática dos principais padrões radiológicos de DPD em imagens de TCAR [7]. O *framework* de software foi integrado a um PACS para reconhecimento de padrões em DPD. Resultados iniciais mostraram que o *framework* desenvolvido proveu um bom potencial para incluir SADM de forma amigável no fluxo de trabalho de radiologia. Esse trabalho foi publicado no *Computer Assisted Radiology and Surgery - Joint Congress of CAR* em 2009 [89].

No domínio de citologia, os nódulos tireoidianos estão entre as doenças mais comuns envolvendo a tireoide. Nódulos palpáveis são encontrados em 4 a 7% da população adulta (com maior incidência entre mulheres e idosos), aumentando para 19 a 67% em exames de ultrassom, devido à sua maior precisão [90]. Embora comuns, apenas 5% dos nódulos são malignos, justificando a necessidade de um diagnóstico mais acurado para início do tratamento mais adequado ao paciente com malignidade. Um SADM capaz de diferenciar entre os tipos de nódulos pode ajudar o patologista a concluir diagnósticos com o fornecimento de informação extra, podendo evitar procedimentos desnecessários em pacientes com nódulos benignos e aumentar a sobrevida dos pacientes no caso de câncer. O uso de sistemas computacionais pode auxiliar o estabelecimento do diagnóstico pela análise de imagens, uma vez que fornecem informações complementares de modo a diminuir o tempo e o esforço necessários para analisar um exame. Tradicionalmente, esses sistemas baseiam-se principalmente em técnicas de Processamento de Imagens e Recuperação de Imagens Baseada em Conteúdo. Segundo pesquisadores, a extração de atributos pode não oferecer ajuda substancial durante o processo de recuperação, pois o profissional de saúde muitas vezes não está familiarizado com os atributos escolhidos, especialmente os de baixo nível como, por exemplo, energia e entropia de uma imagem [91].

Esta pesquisadora acredita que o uso de informações em nível semântico pode melhorar a

precisão da recuperação de imagens, tornando possível ao usuário especificar sua intenção de busca e também aprimorar a interpretação dos resultados. Nesse sentido, esta pesquisadora e colaboradores propuseram o FREDs (*Framework* para Redução da Descontinuidade Semântica em Imagens Médicas) para a definição de mapeamentos conceituais entre o conteúdo de imagens microscópicas e informações textuais presentes em exames.

Os principais requisitos funcionais do FREDs são: (i) permitir que o patologista submeta imagens microscópicas obtidas por meio de microscópios com câmeras digitais; (ii) rotular (identificar e nomear) regiões de interesse em uma imagem médica; (iii) retornar laudos cujas descrições sejam similares a expressões de busca obtidas a partir dos rótulos de uma imagem; (iv) obter e agrupar os diagnósticos de um conjunto de laudos e (v) obter os termos mais frequentes de um conjunto de diagnósticos.

O *framework* FREDs inicial tinha as seguintes cinco interfaces de software: Segmentador, Extrator, Classificador, Recuperador de Informação e Agrupador de Exames. Essas interfaces foram definidas com o objetivo de prover um meio de acesso comum às classes do sistema, permitindo que novos componentes sejam implementados futuramente sem a necessidade de readequar os componentes que os utilizam. A Figura 3.4 apresenta o Diagrama de Classes elaborado, incluindo o diagrama proposto inicialmente para SADM de DPD. O trabalho foi inicialmente publicado no *Congresso de Engenharia Biomédica* (CBEB) em 2008 [7], e posteriormente com extensões no *Workshop de Informática Médica* (WIM) em 2011 [8].

Durante a experimentação do FREDs, a rotulação automática⁴ trouxe rótulos que representavam achados patológicos frequentemente associados a um tipo de câncer de tireoide. Essas palavras-chave dos rótulos permitiram a recuperação de laudos similares às imagens em análise ainda não-diagnosticada. Portanto, uma forma de construir um mapeamento entre imagens médicas e diagnósticos foi com a utilização dos rótulos da imagem. Usando a abordagem proposta foi possível obter termos relacionados a diagnósticos de câncer pela busca em uma base de laudos. O mapeamento criado foi aplicado no contexto de SADM para extrair informações complementares de exames que descrevem componentes celulares similares à aqueles identificados em uma imagem microscópica sendo avaliada pelo patologista.

Em um estudo teórico-prático, esta pesquisadora propôs um método que suporta a transformação de conteúdo das imagens microscópicas de núcleos em informações morfométricas e em características de textura e espectral, quando aplicados a SADM. Sistemas SADM nor-

⁴Usando uma base de exemplos de achados patológicos, a rotulação automatizada foi conduzida por meio da segmentação de uma imagem microscópica; e cada componente segmentado foi submetido à extração de atributos e posteriormente classificado por similaridade quando recebe o rótulo de acordo com sua similaridade com os achados patológicos da mesma base de exemplos.

Figura 3.4: Classes de Software do FREDs. Em cinza, os componentes propostos em [7] e em branco os desenvolvidos em [8].

malmente executam um pipeline com as seguintes tarefas: (i) segmentação, (ii) extração objeto, (iii) rotulação de imagem e (iv) as tarefas classificação. O método proposto estende esse pi-

peline de procedimentos, durante a manipulação e o processamento de imagens, acrescentando análises morfométricas como um método de extração de objeto (caracterização de imagem microscópica). De acordo com os experimentos realizados, o método promoveu um aprimoramento na precisão da classificação automática de imagens microscópicas. O método proposto é mostrado na Figura 3.5. Cada procedimento é detalhado na publicação do trabalho no periódico internacional denominado *Computer, Methods and Programs in Biomedicine* em 2016 [9].

Figura 3.5: Método proposto para um SADM suportado pelo FREDs [9].

3.3.2 Contexto de *Picture Archiving and Communication System* (PACS)

No cuidado com a saúde, o uso crescente de imagens, tais como, raio-X, ultrassons, Resonâncias Magnéticas (RM) e Tomografias Computadorizadas (TC) levou ao desenvolvimento de sistemas computacionais capazes de armazenar e administrar grandes conjuntos de imagens. Com esse objetivo, PACS (*Picture Archiving and Communication System* - Sistema de Comunicação e Arquivamento de Imagens) é um sistema do tipo sistemas baseados em computador para arquivamento, armazenamento, distribuição e acesso conveniente a imagens de diferentes modalidades, os quais devem ser integrados com outros sistemas hospitalares. A utilização de sistemas PACS tem trazido vantagens operacionais e administrativas para as organizações de saúde e de medicina.

O Lyria PACS é um sistema de arquivamento e de comunicação e imagem que corresponde a uma solução completa para o gerenciamento de imagens médicas, em hospitais, centros de imagem e clínicas por profissionais de saúde. Esse sistema oferece acesso às imagens, registros médicos de pacientes, relatórios de diagnóstico, e vários tipos de informações sobre os procedimentos realizados desde o momento da aquisição da imagem até o armazenamento. O Lyria também tem acesso móvel suportado pelas infraestruturas LAN, WAN e Internet. A arquitetura Lyria foi projetada para permitir a escalabilidade, a alta disponibilidade e a resiliência e para

fornecer operação contínua e de qualidade. O Lyria está implantado desde 2011 no Hospital das Clínicas de Ribeirão Preto no campus da USP.

Esta pesquisadora tem sido convidada para colaborar com a equipe de desenvolvimento do Lyria PACS. O primeiro projeto foi a determinação dos benefícios deste sistema, dados os principais desafios enfrentados durante a sua execução, a heterogeneidade de recursos de infraestrutura e da economia monetária obtida após a instalação de Lyria PACS em períodos curtos e médios de tempo. Os resultados desse levantamento foi apresentado no *IEEE International Symposium on Computer-Based Medical Systems* (IEEE CBMS) em 2015 [92]. Para as próximas atividades, planeja-se aplicar as investigações científicas da pesquisadora no contexto relacionamento de informação textual e imagens.

3.4 Relacionamento de Informações Textuais e de Vídeo

Há uma crescente demanda de conhecimento e de manipulação de informação semântica para a personalização, de acordo com os requisitos/características de usuários ou adaptação de informação multimídia. Para realizar o relacionamento automático entre as preferências de usuário e o significado de conteúdos multimídia, por exemplo, técnicas de classificação textual em *closed-caption* em vídeos de programas de notícias para a definição do assunto. Nesse contexto, a técnica de Indexação da Semântica Latente (*Latent Semantic Indexing* - LSI) pode ser explorada para tarefas de classificação de informações para fins de recuperação. Alguns inconvenientes da técnica podem impor limitações, principalmente, quando existe a manipulação de várias coleções. Pelo fato desta pesquisadora ter trabalhado com LSI nos primeiros serviços de criação automática de ligações (ver Capítulo 2), ela foi convidada para comparar os resultados de LSI para relacionar conteúdo textual a *streams* de vídeo com os resultados de Algoritmos Genéticos (*Genetic Algorithms* - GA) para a mesma tarefa.

O LSI é geralmente utilizado para tarefas de recuperação de informação, contudo, nesta colaboração LSI relacionou, por meio de técnicas de classificação, um conjunto de modelos textuais em categorias pré-definidas de vídeos. As informações de texto foram capturadas a partir de fontes de vídeo por meio da análise da *Closed-Caption* (CC) dos mesmos. Para o uso da técnica, houve a remoção de *stopwords* e aplicação de *stemming*. Em seguida, foi gerada a matriz de termos por documentos, a qual foi então decomposta em três novas matrizes usando o método SVD (*Singular Value Decomposition*). Na sequência, um processo de redução de dimensionalidade foi realizado a fim de eliminar ruídos. Finalmente, um conjunto de modelos com palavras-chave para cada categoria de vídeo foi utilizado como instrução de consulta a fim

de classificar os documentos compostos por textos dos CCs, de acordo com as consultas com os termos mais semelhantes. Os modelos (categorias) e algumas palavras-chave comuns para as categorias, respectivamente, são:

- Categoria 1 (Weather & temperature): degree, forecast, rain, sun, frost, hot, cold, clean, fog, maximum, etc;
- Categoria 2 (Politics & government): candidate, congress, constitution, state, federal, mayor, senate, councillor, etc;
- Categoria 3 (Economics & application): bank, capital, deficit, dollar, investment, money, savings, price, flow, etc;
- Categoria 4 (Science & research): result, study, method, paper, scientist, publication, technique, develop, etc;
- Categoria 5 (Sports & medal): player, winner, loser, classification, game, staff, leadership, gold, silver, etc;
- Categoria 6 (Violence & murder): kill, rape, attack, pursue, police, shoot, arrested, armed, hit, steal, escape, etc.

O algoritmo de GA realizou o mesmo pré-processamento com eliminação de *stopwords* e uso de *stemming*, e a montagem da matriz de termos por documentos. Além disso, as duas técnicas compartilharam o mesmo conjunto de modelos, definidos para cada categoria com suas palavras-chave.

Na Figura 3.6, observa-se que o GA apresentou melhores resultados de precisão e de revocação do que LSI, especialmente para as categorias 1, 3, 4, 6 e t (todas as categorias). A Categoria 7 é composta por pedaços de notícias da Categoria 1. A Figura 3.6 mostra que os pedaços de notícias de 1 foram todos classificados com sucesso por GA.

O relacionamento entre informações textuais de *closed-caption* e de vídeos foi realizada pela classificação das informações em categorias pré-definidas e caracterizadas. As técnicas de LSI e de GA utilizaram o elevado nível semântico e de qualidade da informação textual curada para relacioná-la a streams de vídeo. As principais contribuições da investigação foram (i) a descrição de como configurar o algoritmo de LSI padrão e a alternativa método de GA para fins de classificação, utilizando um conjunto de modelos; e (ii) a avaliação de ambas as técnicas para comparação de desempenho. Esta comparação indicou alguns inconvenientes de LSI, o que resultou em um melhor desempenho da alternativa GA. Embora, a técnica de GA tenha

Figura 3.6: Avaliação Geral [10].

apresentado melhores resultados quando comparado com a LSI, ela tem algumas limitações. Este trabalho em colaboração com pesquisadores do ICMC-USP de São Carlos gerou uma publicação apresentada no *Int. Journal Advanced Media and Communication* em 2009 [10].

3.5 Considerações Finais

Apesar da infraestrutura CALiSP ter sido publicada em periódico internacional apenas em 2008, ela foi um trabalho cooperativo com o laboratório Intermídia do ICMC, explorando o resultado do doutorado desta pesquisadora, o CARE (apresentado no Capítulo 2) no contexto de interações multimodais.

Os esforços da pesquisadora com o método de inclusão de informações morfométricas, o FREDs para DPD e para citologia com busca de rótulos em laudos contribuíram com a redução da descontinuidade semântica existente entre a recuperação computadorizada de imagens médicas e a interpretação humana de seu conteúdo, uma vez que o relacionamento (mapeamento) semântico de informações da imagem com outras informações pode resultar na geração de conhecimento. Os trabalhos no contexto multimodal e multimídia com imagens geraram resultados publicados em periódicos internacionais de bom impacto e/ou excelente Qualis na CAPES.

4 Manipulação, Relacionamento e Classificação de Informação Biomédica na Prevenção a Doenças

O sistema de saúde brasileiro é suportado por serviços fornecidos por empresas privadas de saúde e por um sistema de saúde público controlado pelo Estado, o Sistema Único de Saúde (SUS). O SUS é categorizado em três níveis de complexidade: primário ou básico (cuidados preventivos, cuidados primários de saúde e controle de doenças), secundário ou especializado (cuidados de saúde de média complexidade) e terciário ou de alta complexidade. Esse sistema começa concentrando-se em cuidados básicos de saúde, os quais deveriam aprimorar as chances de desempenho da rede de serviços de saúde. Os cuidados básicos de saúde do setor primários devem ser realizados rotineiramente por profissionais de saúde, os quais têm contato direto com os pacientes, bem como com as suas famílias. Assim, os profissionais da saúde podem ser considerados a maior fonte de informações sobre a saúde e o desenvolvimento das famílias brasileiras. Eles são indicados para iniciar cuidados preventivos de saúde. Por exemplo, o papel da equipe de saúde na atenção cuidados primários, principalmente no atendimento pediátrico, é importante porque a equipe deve ser capaz de detectar, o mais cedo possível: (i) as crianças em risco, com o objetivo de promover referências ou intervenções para minimizar ou corrigir problemas detectados e também (ii) as crianças que têm fatores de proteção para a promoção e desenvolvimento da saúde [93]. Usando informações atualizadas, profissionais da atenção primária devem identificar prontamente as crianças que estão em risco e prescrever a intervenção necessária para minimizar futuros problemas de saúde. Este tipo de informação de saúde pode apoiar medidas preventivas de saúde e melhorar o bem-estar dos indivíduos e da sociedade. O conhecimento sobre o processo de desenvolvimento humano é um entendimento muito importante para os profissionais de saúde porque a informação colhida durante o desenvolvimento do paciente pode apoiar iniciativas de prevenção, melhorando a qualidade de vida do paciente.

Ao mesmo tempo, a literatura científica biomédica ou de saúde tem crescido exponencialmente e tornado-se disponível em diferentes repositórios de informações. Por exemplo, o repositório PubMed¹, um repositório de informação livre desenvolvido e mantido pelo Centro Nacional Americano de Informações sobre Biotecnologia (*the U.S. National Center for Biotechnology Information* - NCBI²), é atualmente composto por mais de vinte milhões de citações da literatura biomédica da MEDLINE (*Medical Literature Analysis and Retrieval System Online*). Apesar de os efeitos positivos de grandes quantidades de informação científica, há também alguns aspectos negativos. A grande quantidade de informação científica sobrecarrega os profissionais de saúde interessados em manter-se atualizado. Além disso, as buscas por informações precisas são complexas e demoradas. Outro aspecto negativo é o fato de os usuários manipularem consultas compostas por uma variedade de sinônimos e homônimos para doenças e conceitos de saúde. Além disso, eles devem enviar consultas a diferentes bases de informações e motores de busca. A aplicação de métodos computacionais ou mesmo software para mineração dados ou informações de saúde pode ajudar a melhorar as práticas e os procedimentos de saúde por profissionais da área; por exemplo, a identificação de crianças com problemas de desenvolvimento a partir de informações científicas relacionadas com essas crianças.

Esta pesquisadora tem aplicado suas pesquisas na investigação de relacionamentos entre informações biomédicas. O objetivo é explorar a criação automática de relacionamentos entre documentos das áreas médica e de saúde com conteúdo semelhante para, por exemplo, facilitar e aprimorar estudos de casos clínicos. Para alcançar esse objetivo, esta pesquisadora com seus alunos e colaboradores desenvolveram o Automatic-SL, o CISS, o HSSF, o CISS+, o CISS-SW e o QASF. Esses projetos e suas principais contribuições são apresentados a seguir.

4.1 *Automatic Surveillance Level (Automatic-SL)*

O NIPE (Núcleo Interdisciplinar de Pesquisa e Ensino) realizou estudos com o objetivo de sistematizar os procedimentos para acompanhamento e cuidados de saúde dos pacientes tratados no nível de atenção primária de saúde. Esse grupo propôs uma medida para identificar as necessidades de saúde do paciente, chamada de Grau de Vigilância (GV ou *Surveillance Level* - SL). O GV pode ser usado para informar a recomendação de procedimentos pediátricos em cuidados de saúde primários, identificando fatores de risco significativos e fatores de proteção associados com os pacientes e suas famílias. Diferentes níveis de GV são associados a medidas educacionais e terapêuticas gerais e cuidados especializados [93].

---

¹www.ncbi.nlm.nih.gov/pubmed

²www.ncbi.nlm.nih.gov

Um centro médico da comunidade do bairro Vila Lobato na cidade de Ribeirão Preto é uma unidade de saúde de atenção primária que tem fornecido, ensino e serviços de pesquisa clínica por cerca de 50 anos. O centro médico Vila Lobato tem tradição de atendimento da saúde da criança e do adolescente. A maioria das instituições de cuidados de saúde primários brasileiros presta serviços de saúde apenas para pacientes que apresentam sintomas ou doenças. O centro de Vila Lobato rotineiramente agenda horários para atendimento de saúde preventiva. O GV foi experimentado pela Vila Lobato para apoiar os seus programa de cuidados da saúde e serviços relacionados. Um dos objetivos da instituição era promover uma estreita relação entre as equipes multidisciplinares de saúde.

No contexto dos programas de cuidados da saúde do centro médico da Vila Lobato, o PRO-ESPA (Programa de Educação em Saúde para Pais e Famílias) sugere ações de intervenção na atenção básica a fim de tornar pais e familiares aptos para a promoção de um desenvolvimento saudável de seus filhos e identificação de alterações no desenvolvimento normal. Na idealização do PROESPA, os pais e familiares alvos das ações do programa são principalmente aqueles que possuem bebês recém-nascidos, acompanhando o desenvolvimento desde o início da vida [94]. A equipe de profissionais do PROESPA é interdisciplinar, sendo composta por psicólogos, fonoaudiólogos, fisioterapeutas, assistentes sociais, pedagogos e terapeutas ocupacionais, além de médicos e enfermeiros que podem vir a encaminhar um indivíduo para o PROESPA. Os funcionários e profissionais do PROESPA são continuamente treinados para a capacitação de atribuição do GV. Há também uma proposta de que crianças em idade escolar possam ser encaminhadas para o PROESPA a partir do PROISE (Programa de Integração Saúde e Educação), ou seja, encaminhadas a partir das escolas.

No cenário desses dois programas na Vila Lobato, a pesquisadora e seu grupo de trabalho começaram suas atividades em 2004. Para identificar as crianças com problemas de desenvolvimento, esta pesquisadora e seus alunos definiram uma arquitetura que apoia a geração automática de graus de vigilância (GV) para crianças atendidas por profissionais da saúde com a intenção de auxiliar na tomada de decisões [11]. Usando classificadores de aprendizagem de máquina, atribui-se automaticamente os níveis de vigilância para pacientes com base em informações do paciente, após cada consulta médica nos centros de cuidados de saúde primários. A atribuição manual de GV é uma tarefa trabalhosa que exige pessoas treinadas e avaliação personalizada. Graus diferentes de GV estão associados com medidas educativas e terapêuticas gerais e especializados, de acordo com a seguinte escala [95]:

- GV-Rotina: medidas de rotina por parte dos estabelecimentos de saúde primários;
- GV-1: medidas educativas;

- GV-2: medidas educativas e terapêuticas;
- GV-3: forte necessidade de medidas terapêuticas, educativas e especializadas;
- GV-Urgência: paciente necessita de cuidados urgentes.

O primeiro resultado de colaboração foi o desenvolvimento de um protótipo de sistema de registro eletrônico dos pacientes no contexto das disciplinas de Engenharia de Software, de Programação Orientada a Objetos e de Banco de Dados. As disciplinas de Engenharia de Software e de Programação Orientada a Objetos estavam sendo ministradas por esta pesquisadora. O trabalho desenvolvido nesse protótipo para o CMSCVL foi redigido como poster e foi apresentado no 13º Simpósio Internacional de Iniciação Científica da USP (SIICUSP), em novembro de 2005 [62].

Nos anos seguintes, nas mesmas disciplinas os alunos desenvolveram uma nova versão do sistema. No contexto de trabalhos de conclusão de curso, duas alunas orientadas pela pesquisadora deram continuidade a parceria com o Centro Médico da Vila Lobato. A aluna Daniane de Paula criou um sistema de informação para suportar as atividades do PROESPA e relacioná-las com o registro médico do paciente[96]. A aluna Juliana Pollettini propôs a definição automática, ou semi-automática em alguns casos especiais, do GV do paciente a partir da manipulação e análise de informações de prontuários de atendimento usando RF (*Relevance Feedback*) e DRG (*Diagnosis Related Groups*) [97]. Com apoio da FAPESP com uma bolsa de iniciação, o sistema de definição GV, posteriormente chamado de Automatic-SL, foi aprimorado.

A Figura 4.1 apresenta a arquitetura do Automatic-SL, a qual foi organizada em três camadas: (i) Camada de Apresentação (ver Figura 4.1a); (ii) Camada de Classificação, uma plataforma de software suportada por seis módulos (cinco módulos de classificação e um módulo de análise linguística responsável pelo pré-processamento de linguagem) (ver Figura 4.1b); e (iii) Camada de Armazenamento, uma camada que processa a informação com base no histórico médico do paciente e permite a inclusão das informações do paciente (ver Figura 4.1c).

A Camada de Apresentação (ver Figura 4.1a) é a interface gráfica do usuário que suporta (i) recomendação do GV como uma segunda opinião para profissionais de saúde (esta função pode ser usada, por exemplo, durante as consultas médicas); (ii) atualização do GV; (iii) mapa de projeções georeferenciadas de GV; (iv) reavaliação das recomendações GV-3 ou GV-emergência; (v) a configuração do software (por exemplo, os parâmetros dos classificadores); e (vi) representação visual das informações geradas pelos classificadores, incluindo ordenação por precisão dos classificadores. Após as informações do paciente serem inseridas, a camada de classificação atribui o GV para o paciente. O profissional de saúde também pode acessar

Figura 4.1: (1) Arquitetura do sistema Automatic-SL organizada em três camadas: a) Camada de Apresentação, b) Camada de Classificação, e c) Camada de Armazenamento. A Camada de Classificação tem cinco módulos de classificação: KNN-SL, ANN-SL, RF-SL, DT-SL e vote-SL e um módulo de linguística [11].

um mapa de projeção de informações georreferenciadas com base em endereços de pacientes com GV para análise médica ou administrativa. O aplicativo Google Maps foi utilizado para gerar o mapa de projeção de endereços para todos os pacientes recentemente atribuídos um nível de GV-3. O SLS são georreferenciados com base em endereços de pacientes no sistema de registro médico desenvolvido em 2004 pelos alunos de Informática Biomédica nas disciplinas do curso. O mapa de projeção do Automatic-SL pode ativar a vigilância, por exemplo, de surtos epidêmicos pelas instituições de saúde e governamentais. O algoritmo de apresentação pode ser usada para estabelecer o mapa de qualquer tipo de GV. O mapa de GV foi publicado e apresentada no Congresso Brasileiro de Informática em Saúde em 2010 [98].

Na Camada de Classificação (ver Figura 4.1b), a primeira versão do Automatic-SL aplicou classificadores do tipo K-vizinhos mais próximos (KNN-SL) e Redes Neurais Artificiais (ANN-SL) [99]. A segunda versão incluiu realimentação de relevância (RF-SL) [100]. Finalmente, a terceira versão possuiu suporte de um módulo linguístico, estratégias de Árvore de Decisão (DT-SL) e um classificador de voto majoritário (Vote-SL) [100]. Usando UMLS³, o Automatic-SL foi remodelado para (i) a definição automática de medidas que identificam pessoas requerendo diferentes GVs [102]; (ii) a determinação e análise de tipos semânticos

³O *Unified Medical Language System* (UMLS), mantido pela Biblioteca Nacional de Medicina dos Estados Unidos, é uma importante fonte de informação que contribui para o processamento e gestão de documentos biomédicos. O UMLS compreende uma meta-enciclopédia, uma rede semântica, e um léxico especializado. Este conjunto de tecnologias e fontes de conhecimento foi concebido para ser usado por uma variedade de aplicações, minimizando o problema de diferentes maneiras em que um conceito pode ser expresso em fontes de informação biomédicas [101].

para aprimorar a determinação automática de GV's [103]; e (iii) a recomendação de Graus de Vigilância para Pacientes do Setor Primário [104]. A cada extensão do Automatic-SL, novos desafios geraram novas investigações e resultados publicados em diferentes congressos e simpósios nacionais [100, 102, 104, 98, 104] e internacionais [99], e no periódico internacional *Journal of Medical Systems* (JMS) em 2012 [11], como foi apresentado nesta seção.

4.2 *Chronic Illness Surveillance System (CISS)*

Doenças crônicas são consideradas sérios problemas de saúde e a principal causa de mortalidade entre homens e mulheres (60% de todas as mortes) — como observado pela Organização Mundial de Saúde [105]. Essas doenças têm etiologias multifatoriais causadas pela interação de vários fatores comuns, os quais incluem genes, nutrição e estilo de vida [106]. Uma dieta pouco saudável, sedentarismo e uso de tabaco são os principais fatores de risco que contribuem para o aparecimento dessas doenças. Pelo menos 80% de todas as doenças do coração, casos de AVC e diabetes tipo 2 poderiam ser evitados se esses grandes riscos fossem eliminados — como também observa a Organização Mundial de Saúde [107] [108].

A medicina genômica sugere que a exposição a fatores de risco durante a infância (no momento da concepção, durante a vida fetal, a primeira infância e a infância) pode influenciar a expressão do gene e, conseqüentemente, induzir o desenvolvimento de doenças crônicas na idade adulta [109], uma vez que esses fatores podem induzir variações em processos de expressão do gene. Considerando que a interação de genes e fatores ambientais desempenham um papel importante em doenças complexas, é possível sugerir uma relação entre a exposição precoce aos fatores de risco e doenças crônicas em adultos. Na década de 90, Barker defendeu a necessidade de um estudo mais aprofundado sobre epigenética⁴, em uma tentativa de prevenir doenças de alta prevalência, como doenças cardiovasculares, diabetes e obesidade [111][112]. David Barker menciona, por exemplo, que muitos bebês e fetos humanos são obrigados a se adaptar a uma oferta limitada de nutrientes, o que influencia mudanças permanentes em sua fisiologia e metabolismo. Essas mudanças podem dar origem a uma série de doenças na vida adulta, como diabetes e doenças coronárias [109]. De acordo com Wadhwa e outros, a “Hipótese de Barker” surgiu há aproximadamente 25 anos e estimulou o interesse de pesquisadores pelas origens fetais das doenças dos adultos, levando à formação de uma sociedade internacional para estudar o tema [113]. A *International Society for Developmental Origins of Health and Disease* (DOHaD) concentra seus estudos nos primeiros estágios do desenvolvimento humano e prevê um novo paradigma de prevenção para doenças crônicas comuns [114].

⁴Estudo de características herdáveis do genitor para a prole que não resultam de uma mutação no DNA [110]

Estudos científicos geram conhecimento que pode ser utilizado para benefício da população. Artigos científicos com descobertas e estudos inovadores no domínio de atuação da DOHaD indicam que a epigenética deve ser explorada para prevenir doenças de alta prevalência (como doenças cardiovasculares, diabetes e obesidade). Essas informações normalmente estão apresentadas em vastos repositórios de artigos científicos, que podem ser consultados por profissionais de diversas áreas. A grande quantidade de artigos, todavia, dificulta a atualização de profissionais, uma vez que as buscas por artigos científicos com a informação exata que se deseja tornam-se complexas e dispendiosas em relação ao tempo gasto na procura e análise dos resultados. Algumas tecnologias e técnicas computacionais podem apoiar o gerenciamento dos grandes repositórios de informações biomédicas e a geração de conhecimento, por exemplo, pelo relacionamento de resultados de pesquisas de diferentes pesquisadores, institutos ou até mesmo áreas. As técnicas de processamento semântico e mapeamento de documentos para estruturas conceituais, por exemplo, podem ser úteis para a identificação de conceitos e de relações e, conseqüentemente, para facilitar a descoberta do conhecimento. Tecnologias como essas podem, portanto, ser usadas para facilitar a prática da medicina atual, a qual é apoiada pela descoberta de evidências.

Para associar crianças com problemas de desenvolvimento e a literatura científica e alertar profissionais da saúde sobre os problemas de saúde, o grupo desta pesquisadora desenvolveu o CISS (*Chronic Illness Surveillance System*). O CISS pretende prover aos profissionais de saúde informações a cerca de exposições a fatores ambientais no princípio da vida, os quais podem induzir a modificações no desenvolvimento humano, podendo gerar impacto na saúde na vida adulta e causar risco de doença [12]. Por exemplo, os profissionais da saúde devem estar cientes de que, grávidas com sobrepeso expõem seus fetos a um fornecimento desequilibrado de nutrientes com quantidades excessivas de açúcar. Conseqüentemente há uma tendência para a subnutrição pré-natal com crescimento retardado seguido por uma melhor nutrição pós-natal com o crescimento acelerado [115]. Outros problemas associados à exposição ambiental são diabetes tipo 2 e distúrbios metabólicos, osteoporose, doença pulmonar obstrutiva crônica, algumas formas de câncer e algumas doenças mentais [116]. Para favorecer boas condições de vida na idade adulta, é essencial compartilhar informação e alertar os profissionais de saúde. Com o objetivo de divulgar informações preventivas, o serviço CISS associa fatores de risco genéticos e epigenéticos em termos de doenças crônicas apresentados em artigos científicos com os registros clínicos de pacientes. Ao estarem cientes dessa abordagem, os profissionais da saúde podem criar uma rotina clínica com as famílias buscando melhores condições de crescimento. O CISS utiliza termos ontológicos, em vez de palavras-chave para compor uma coleção mais precisa de artigos científicos em epigenética a serem relacionados com casos clínicos.

A Figura 4.2 apresenta a arquitetura do sistema com módulos e relacionamentos. O CISS tem um módulo chamado *Collection Creation/Update* (Figura 4.2(a)) que, usando uma interface de pesquisa (Figura 4.2(b)), é responsável por interações com PubMed (Figura 4.2(c)) - o módulo rotineiramente procura e recupera trabalhos usando a interface que explora as APIs *Entrez Programming Utilities* e *Biopython Project API*, com conceitos de ontologias sobre o domínio fatores de risco genéticos e epigenéticos (Figura 4.2(d)). As consultas submetidas ao PubMed são compostas por termos da CDO (*Chronic Disease Ontology*)⁵ e do UMLS (*Unified Medical Language System*) para recuperar uma coleção de documentos a partir do repositório público (Figura 4.2(e)). Os artigos são textualmente processados (Figura 4.2(f)-(g)) e, em seguida, armazenados para uso futuro em um banco de dados local (Figura 4.2(h)). Este processo (Figura 4.2(a)-(h)) é repetido rotineiramente para que os novos artigos incluídos na coleção pública sejam textualmente processados e os resultados correspondentes sejam armazenados no banco de dados. O principal objetivo é compor uma coleção de artigos científicos pré-processados para apoiar tarefas de recuperação de informação (Figura 4.2(h)).

Para aumentar a especificidade de termos, bem como relacionar conceitos em inglês e em português, o módulo *Concept Recognition* (Figura 4.2(p)) procura conceitos do UMLS, um exemplo de um recurso linguístico (Figura 4.2(o)). O uso de UMLS foi essencial para o trabalho porque: (i) ele suporta a composição da consulta executada pelo módulo *Collection Creation/Update* (Figura 4.2(a)); (ii) ele ajuda a diminuir o número de termos considerados pelo processamento de similaridade entre os documentos (Figura 4.2(k)); (iii) ele ajuda a aumentar a especificidade de termos manipulados; e (iv) ele relaciona termos entre inglês e português.

Para associar artigos científicos a registros clínicos (Figura 4.2(i)) apresentados por um profissional de saúde, CISS processa o registro clínico (Figura 4.2(g) e Figura 4.2(j)) e, em seguida, chama o módulo responsável pelo processamento da similaridade entre documentos (Figura 4.2(k)). Esse módulo acessa (Figura 4.2(l)) a versão pré-processada da coleção de artigos científicos (Figura 4.2(h)) e os artigos com os mais altos valores de similaridade com os registros clínicos são recuperados. Os artigos selecionados são apresentados aos profissionais de saúde em uma interface gráfica do usuário (GUI) com opção para visualizar uma lista de fatores de risco (Figura 4.2(n)).

O processamento de texto, apoiado por recursos linguísticos (Figura 4.2(o)), inclui a remoção de *stopwords*, processamento de n-gram, o reconhecimento de conceitos (Figura 4.2(p)) e o cálculo de pesos para os conceitos. CISS usa a estratégia *tf-idf*, frequências de termos e

⁵Verma et al. [106] desenvolveram a CDO para armazenar, reutilizar e descobrir novos conhecimentos a partir três tipos de doenças crônicas: doenças cardiovasculares, diabetes tipo 2 e obesidade. A CDO inclui informações sobre as relações entre genes e mutações, bem como os dados do histórico de saúde, nutrição e vida.

Figura 4.2: Módulos de processamento e de armazenamento da infraestrutura CISS. Uma *Graphical User Interface* (GUI) pode ser apresentada aos pediatras para ilustrar a recuperação de artigos científicos em inglês após a consulta implícita considerando um registro médico em português [12].

frequências inversas de documentos, para atribuição de pesos aos termos de índice (conceitos) dos documentos. O processamento de registros clínicos é semelhante ao tratamento de documentos do acervo de trabalhos científicos. A diferença entre os dois é que, para os registros clínicos em português, a remoção de *stopwords* é realizada usando a lista de *stopwords* em português, a partir do projeto Snowball [117]. O processamento dos registros clínicos também difere do anterior na medida em que, após o processamento de n-gram e a identificação dos conceitos de saúde associados com o n-gram, um vetor contendo os restantes conceitos é construído e submetido ao processo de cálculo de similaridade. Esse processo identifica as relações entre registros clínicos e artigos da coleção.

A matriz de termos de índice do Automatic-SL é semelhante a matriz de conceitos do CISS. A aplicação do reconhecimento de conceitos e a criação da matriz de pesos, de acordo com estes conceitos reconhecidos, são perfeitamente compatíveis com o Automatic-SL. O CISS e sua infraestrutura foram publicados, respectivamente em congressos nacionais e internacional [118, 119, 120] e no periódico internacional *BMC Medical Genomics* [12].

Com base em abstrações dos sistemas Automatic-SL e CISS, foi proposto o HSSF, apresentado a seguir. Posteriormente, novos recursos foram adicionados ao CISS e o HSSF foi atualizado para se tornar mais abrangente e genérico. O HSSF e esses novos recursos, os quais geraram dois novos serviços – o CISS+ e o CISS-SW, são apresentados na próxima seção.

4.3 Health Surveillance Systems Framework (HSSF)

Esta pesquisadora definiu um *framework* para apoiar o desenvolvimento de sistemas de vigilância em saúde. O HSSF foi construído por meio de abstrações e de generalizações dos softwares de vigilância desenvolvidos pela pesquisadora e seu grupo de pesquisa, o Automatic-SL e o sistema CISS. O intuito foi o de tornar o esforço despendido no desenvolvimento desses sistemas reutilizável em termos de projeto e de funcionalidades.

Como ponto de partida, HSSF foi projetado com suporte das classes de software do sistema de vigilância CISS, com incrementos de classificadores das versões do Automatic-SL. Roberts e Johnson sugerem um número pequeno e finito (por volta de três) de exemplos para criar um *framework*, de modo a torná-lo suficientemente genérico [53].

Os sistemas CISS e Automatic-SL foram abstraídos a partir de suas classes de software coincidentes, de seus objetivos e de suas colaborações. Por exemplo, o classificador de *relevance feedback* (RF-SL) do Automatic-SL gera informações estruturadas, a partir de registros médicos como *bags of words*, realizando também a eliminação de *stopwords* e a geração de *stemming* com o objetivo de gerar uma matriz de termos com pesos. Esta matriz resultante é semelhante à matriz conceito utilizada pelo CISS para calcular as similaridades entre artigos científicos e registros médicos. A diferença é que a matriz do CISS utiliza conceitos UMLS, em vez de termos simples da matriz do Automatic-SL. No entanto, a aplicação de reconhecimento conceitos e a construção da matriz de pesos são perfeitamente aplicáveis ao Automatic-SL, e não apenas ao módulo de RF-SL.

A seguir, o HSSF é detalhado em termos de infraestrutura, diagrama de classes e casos de uso. O HSSF está em processo de aceitação (submetido com *Minor Revisions*) pelo *Journal of Biomedical Informatics* [13]. Seu diagrama foi submetido para o *ACM DocEng 2016* [14].

4.3.1 Infraestrutura de Software

Considerando os objetivos do Automatic-SL [11] e do CISS [12], o HSSF suporta os seguintes três casos de uso e seus requisitos: (i) criação de uma coleção de artigos científicos de

um domínio específico; (ii) recuperação de artigos científicos associados a registros clínicos; e (iii) classificação de pacientes em grupos de risco.

A Figura 4.3 apresenta a infraestrutura do HSSF dividida nas Camadas de Apresentação (*Presentation Layer*), de Lógica (*Logic Layer*) e de Armazenamento (*Storage Layer*). Cada camada tem seus módulos de processamento e de *programming utilities*. A Camada de Lógica é uma estrutura de software composta de classes abstratas do domínio (a *Business Layer*) e de pacotes externos, e duas camadas de conectores, a Camada de Comunicação (*Communication Layer*). A Camada de Comunicação fornece os meios de comunicação necessários (tais como ontologias e conectores de banco de dados) para as camadas de apresentação e de armazenamento.

A Camada de Apresentação apresenta diferentes *views* e *templates* como GUI para permitir o acesso principalmente por dois tipos de usuários: (1) os profissionais de saúde, que podem analisar grupos de risco classificados automaticamente usando SL ou receber recomendações de trabalhos relacionados com prontuário do paciente durante uma consulta médica, e (2) os pesquisadores que estão interessados em investigar a relação entre fatores de risco, doenças crônicas e o desenvolvimento dos pacientes ou de grupos de risco. As duas camadas de comunicação são basicamente compostas por conectores para ferramentas, ontologias e fontes de conhecimento. A camada superior conecta a informação resultante dos dois tipos de usuários, por meio de interfaces gráficas de usuário. A camada de comunicação inferior integra a Camada de Negócios com novos recursos estendidos por ferramentas e fontes de conhecimento. Essa camada é também composta por módulos de comunicação com bases de dados e de ontologias da camada de armazenamento. A Camada de Negócios é composta principalmente pelos módulos do CISS e o módulo de classificação do Automatic-SL.

Na Camada de Negócios, o módulo *Search For Papers* interage com os repositórios públicos de artigos científicos. Esse módulo coleta e atualiza a coleção de artigos. Atualmente, o *crawler* de repositórios utiliza conceitos de ontologias de domínio para se focar em assuntos de interesse. Para registros clínicos, não há *crawler*, porque todos os registros clínicos são considerados como associados a artigos científicos.

O módulo de *Textual Processing* é composto de *programming utilities* e módulos para *Papers Processing*, *Clinical Records Processing* e *Natural Language Processing*. O módulo de *Textual Processing* processa informação textual de um conjunto de registros médicos e da coleção de artigos científicos selecionados, ambos armazenados na Camada de Armazenamento. Cada documento (registro clínico ou artigo) é processado para identificar termos simples e complexos pelos módulos *Papers Processing* e *Clinical Records Processing*. O módulo

Figura 4.3: Infraestrutura do HSSF composta por módulos de processamento e de armazenamento [13].

Natural Language Processing aplica mecanismos de processamento de linguagem natural, tais como processamento *n*-grams, *stemming*, remoção de *stopwords* e reconhecimento de termos simples e compostos. O processamento textual do HSSF, suportado por recursos linguísticos, inclui a remoção de *stopwords*, o processamento de *n*-grams, o reconhecimento de conceitos e os cálculos de peso para cada conceito. A remoção de *stopword* para a coleção de artigos e a coleção de registros clínicos são respectivamente baseados na lista de *stopword* de *programming utilities* tal como Snowball⁶. O processamento *n*-grams usa um conjunto de módulo de código aberto Python NLTK, dados linguísticos e documentação para pesquisa e desenvolvimento em processamento de texto e análise de linguagem natural [121]. Os termos reconhecidos são atribuídos pesos estatísticos e, em seguida, eles são armazenados.

Recursos linguísticos são manipulados pelo módulo de *Concept Recognition* e eles suportam a associação entre diferentes conceitos lexicais. Registros clínicos podem ser manipulados em um idioma específico e artigos podem ser processados em outro idioma. Por exemplo, o módulo *Concept Recognition* explora classes e métodos do UMLS para identificar conceitos de tópicos relacionados a saúde, especialmente epigenética.

O processamento de registros clínicos não é idêntico ao de artigos porque após o processamento de *n*-gram e a identificação de conceitos, uma matriz de consulta contendo os conceitos

⁶snowball.tartarus.org

restantes é construída para registros clínicos, enquanto uma matriz de pesos é composta pelos artigos científicos. Ambas matrizes são submetidas ao módulo *Similarity* responsável pelos cálculos de similaridade. A infraestrutura do HSSF permite que o módulo *Similarity* calcule medidas de similaridade entre os documentos e os registros clínicos e também ele é capaz de aplicar *relevance feedback*, de forma automática. Registros clínicos são manipulados na língua nativa do profissional da saúde. A infraestrutura do HSSF foi submetida em 12 de maio de 2016 está em *Minor Revision* pelo JBI [13].

4.3.2 Diagrama de Classes e Pontos de Reúso

Roberts e Johnson indicam que um *framework* de software é usualmente composto por pontos flexíveis ou adaptáveis (*hot spots*), os quais podem ser modificados de aplicação a aplicação e pontos fixos (*frozen spots*), os quais não podem ser facilmente modificados para cada aplicação [53].

4.3.2.1 Identificação de Pontos Flexíveis e de Pontos Fixos

O HSSF provê pontos adaptáveis que facilmente permitem: (i) a inserção de outros repositórios científicos de informação além do PubMed; (ii) o uso de ontologias de outros domínios para criar consultas com o intuito de filtrar artigos científicos do repositório de informação; (iii) a exploração de outras ontologias e/ou dicionários de domínio, por exemplo, para o reconhecimento de termos médico e biomédicos além do UMLS; (iv) a manipulação de diferentes tipos de registros clínicos ou outros documentos do domínio de saúde; e (v) o uso de outras técnicas de classificação e categorização. Considerando esses pontos flexíveis, o menos flexível é a exploração de outras ontologias ou dicionários para reconhecimento de conceitos. O HSSF realiza processamento multilíngue (atualmente entre inglês e português), consequentemente, é necessário aplicar recursos linguísticos para relacionar conceitos multilíngues e definir relacionamentos semânticos. No domínio de saúde, UMLS ainda representa a melhor opção para reconhecimento de conceitos devido à sua abrangência e qualidade. Entretanto, uma extensão do *framework* HSSF para manipular outros recursos linguísticos além do UMLS, já está projetado no HSSF. Os pontos adaptáveis citados e as interfaces de usuários são as características de flexibilidade mais importantes do HSSF.

Como pontos fixos, pode-se mencionar as classes abstratas que suportam os pontos flexíveis. Por exemplo, uma classe abstrata para ilustrar atributos comuns de diferentes grupos de documentos e classes abstratas para representar atributos de diferentes tipos de documentos. Outros pontos fixos consistem de classes responsáveis por identificar relacionamentos entre documen-

tos e realizar processamento textual (eliminação de *stopwords* e processamento de *n*-grams). Uma análise dos pontos fixos e flexíveis permite perceber que os pontos fixos suportam o método proposto de relacionamento de informação, uma vez que os pontos flexíveis são as classes abstratas que realizam conexão com mecanismos computacionais externos.

4.3.2.2 Diagrama de Classes

A Figura 4.4 ilustra o diagrama de classes do *framework* HSSF. Este diagrama apresenta as classes responsáveis: (i) pela manipulação de um conjunto de artigos científicos como as classes *CreateUpdateCollection* e *Collection*, as quais buscam e fazem *download* de documentos para compor a coleção científica; (ii) pelo processamento textual de documentos, tais como as classes *Document* e *Term*; (iii) pela identificação de conceitos de saúde como *Concept*, *NGramsProcessing* e *MeshTerm*; (iv) pela medição de similaridade entre registros clínicos e documentos, como *RelationshipIdentification*; (v) pela classificação dos pacientes em grupos de risco, de acordo com doenças ou nível de vigilância como *PatientsClassifier*; e (vi) pelo estabelecimento de comunicação entre as Camadas de Negócio e de Apresentação *PatientsClassification*, *PaperRetrieval* e *Pediatrician*. Cada tipo de responsabilidade realizada por um grupo de classes foi colorido por uma cor diferente. Pacotes externos são de cor marrom. Eles foram incluídos como exemplos de instâncias, mas eles são pontos flexíveis para serem alterados por qualquer aplicação, que esteja instanciando o HSSF.

As classes de cor azul correspondem ao módulo *Papers search*. A classe *CreateUpdateCollection* interage com, por exemplo, os pacotes *Biopython* e *Entrez Programming Utilities* para buscar e fazer *download* de artigos de repositórios online. Essas classes têm também métodos auxiliares para manipular ontologias como o pacote da CDO. As classes *PubmedCollection* e *Collection* iniciam o processamento de uma nova coleção de documentos *downloaded* e salva esses documentos na base de dados. A classe *Collection* é abstrata e a classes *PubmedCollection* estende-a. As classes pertencentes ao módulo *Textual Processing* estão apresentadas em tons de verde. Cada tom representa o tipo de informação textual manipulada pela classe (*ScientificPaper*, *ClinicalRecord* e ambos). A classe abstrata *Document* tem métodos responsáveis por fazer o processamento geral (*n*-grams, operações com strings, e manipulação de esquemas de atribuição de pesos) para os tipos de documentos: artigos científicos e registros clínicos. As classes *ScientificPaper* e *ClinicalRecord* têm métodos responsáveis por processamentos específicos. Por exemplo, a *ScientificPaper* possui métodos para processar títulos de artigos e termos de dicionários. Por outro lado, a *ClinicalRecord* tem métodos para processar registros e formular consultas com propósito de recuperação de informação. Além disso, a classe

ClinicalRecord ativa a classe *RelationshipIdentification* e recupera os artigos para um registro clínico em avaliação. A classe *RelationshipIdentification*, do módulo *Similarity* representada em cor vermelha, é constituída de métodos para criar a matriz de pesos, aplicar *relevance feedback*, calcular similaridades entre documentos e ordenar os documentos mais similares a uma consulta. As classes amarelas compõem o módulo NLP (*Natural Language Processing*). Esse módulo inclui, além de *NGramsProcessing* e *MeshTerm*, as classes *Concept* e *Concept-FromUMLS*, as quais interagem com UMLS para encontrar conceitos relacionados a termos de documentos. Essas classes acessam o módulo CR (*Concept Recognition*) composto por pacotes. Essa interação foi desenvolvida em Java e é capaz de identificar conceitos da área de saúde a partir, por exemplo, da API do serviço Web UTS (*UMLS Terminology Service*) ou da ferramenta MetaMap [122]. A primeira prova de conceito do HSSF, CISS+, experimentou essas duas instâncias de pacotes. Os pacotes podem estender o significado semântico de cada termo extraído dos artigos científicos e de registros clínicos.

A cor roxa está ilustrando o módulo *Classification*, com classes estendendo os algoritmos de aprendizado da Weka e do classificador *RFCat Classifier*, desenvolvido para o módulo RF-SL do Automatic-SL. Finalmente, as classes em cor branca fazem parte das Camadas de Comunicação e de Apresentação (GUI). Os pontos flexíveis e fixos, e o diagrama de classes foram apresentados em um artigo submetido ao *ACM DocEng 2016* que ocorrerá em setembro deste ano.

Esta pesquisadora considera que o processamento para reconhecimento de conceitos foi um módulo fundamental em um sistema de extração de informação quando resultados precisos são desejados. A seguir, provas de conceito, explorando a infraestrutura e o diagrama de classes do HSSF são apresentadas. A primeira prova de conceito que instanciou o HSSF, foi o CISS+. O CISS+ diferenciou o uso dos pacotes UTS e MetaMap como duas abordagens para investigar a hipótese sobre a relevância do reconhecimento de conceitos no processamento textual de um domínio específico como o domínio médico ou, especificamente, epigenético.

4.3.3 New Chronic Illness Surveillance System (CISS+)

Para expandir o significado semântico dos termos de artigos científicos e de registros clínicos, o CISS+ foi criado com base: (i) em um cuidadoso estudo comparativo do uso do UTS do UMLS e da ferramenta MetaMap [122]; e (ii) na aplicação de técnicas de *relevance feedback*, principalmente expansão automática de consultas [30]. O desenvolvimento do CISS+ foi perfeitamente suportado pelo HSSF, apenas demandou o uso de dois novos pacotes externos (UTS e MetaMap) pela classe *ConceptRecognition* e a manipulação de novos métodos na classes *Si-*

ilarity (ver Figura 4.4). A ideia era identificar conceitos mais precisos. Por essa razão, foram realizadas buscas na literatura por técnicas de *relevance feedback* e por sistemas capazes de mapear textos livres de artigos em inglês para conceitos ou termos de ontologias ou dicionários.

Em relação aos sistemas de mapeamento de texto, as ferramentas Mgrep, Torch, MetaMap e UTS foram investigadas. Porém, a MetaMap e UTS atendiam os requisitos de uso.

Considerando a Camada Lógica de Negócios da Figura 4.3, o CISS+ reusou todos os módulos de processamento. Os novos recursos linguísticos (pacotes) foram ativados pelo *Concept Recognition* para melhor manipulação da identificação de conceitos de epigenética devido à complexidade da linguagem desse domínio. Consequentemente, a classe *ConceptRecognition* executou métodos dos pacotes UMLS-UTS e UMLS-MetaMap⁷. Usando UTS da Camada de Comunicação, CISS+ manipulou termos resultantes do processamento *n*-gram no módulo *Textual Processing*. O módulo *Concept Recognition* foi capaz de reconhecer conceitos de cuidados à saúde, por meio do UTS, e relacioná-los aos termos resultantes do processamento *n*-gram de artigos e de registros clínicos. Usando MetaMap, o módulo *Concept Recognition* interagiu com uma API para acessar o módulo *Scheduler*, submetendo novos artigos científicos *in batch*. Para UTS e MetaMap, os módulos do CISS+ foram desenvolvidos em Java e Python, e foram conectados usando Python Py4J.

Em termos do uso de *relevance feedback*, o módulo *Similarity* calcula medidas de similaridade entre os artigos e os registros e também executa o mecanismo automático de *relevance feedback*. Para recuperar documentos mais relevantes de uma coleção, consultas de usuário podem ser seletivamente modificadas pela aplicação de expansão de consultas [57][19]. Para aplicar expansão de consulta, informações do conjunto de artigos científicos foram consideradas. Com esse objetivo, CISS+ implementou três abordagens explorando a formulação de Rocchio [57] recriar os vetores de consulta.

A primeira expansão de consulta utilizou meta-informação dos “*MeSH Terms*” dos artigos científicos do PubMed. Essa abordagem considerou o conjunto de artigos contendo termos MeSH Terms relacionados a consulta (um registro médico) como documentos relevantes a serem aplicados na formulação de Rocchio. Essa decisão foi tomada porque os termos MeSH são atribuídos a um artigo por especialistas do domínio da Biblioteca Nacional de Medicina (*National Library of Medicine*) nos EUA. Esses especialistas analisam os assuntos de cada artigo e somente termos MeSH relevantes e específicos são atribuídos. Usualmente, cada artigo contém

---

⁷UMLS distribui terminologia chave, classificação, padrões de codificação e recursos associados para promover a criação de sistemas e serviços de informação biomédica mais eficazes e interoperáveis [101]. UTS fornece acesso à fontes de conhecimento e as ferramentas UMLS. MetaMap é um programa que dá acesso aos conceitos UMLS no Metathesaurus

entre dez e doze termos MeSH. Em uma segunda tentativa de expansão, foram considerados como documentos relevantes o conjunto de documentos que tinha conceitos identificados depois do processamento n -gram dos registros médicos. Se um registro médico tem pelo menos um conceito maior que 1-gram, esse é um documento relevante. Documentos não relevantes têm somente conceitos identificados por 1-gram. Finalmente, a última tentativa aplicou expansão de consulta considerando meta-informação do campo “*Publication Type*” dos artigos PubMed. Portanto, documentos relevantes foram compostos por valores “*review*”, “*editorial*” e “*comment*” atribuídos ao atributo “*Publication Type*”. De acordo com os especialistas em epigenética, artigos e revisões de estudos clínicos e epidemiológicos devem ter um maior valor para o CISS+, porque este tipo de artigo descreve informações importantes da medicina experimental em termos de alterações genéticas raras, que normalmente não despertam o interesse durante a prática clínica. Por exemplo, esses especialistas comentaram que a síndrome de Prader-Willi é um problema complexo e raro relacionado a alterações epigenéticas com um tipo diferente de acometimento, como origem fetal da síndrome metabólica. Um artigo de investigação sobre a síndrome de Prader-Willi apareceu nos experimentos do CISS+.

Os resultados iniciais do CISS+ foram redigidos em um artigo publicado no congresso internacional *IEEE Computer Based Medical Systems* (CBMS) em 2015 [30].

4.3.4 *Chronic Illness Surveillance System by Semantic Web* (CISS-SW)

O CISS-SW é o sistema CISS apoiado pela Web Semântica (*Semantic Web*) com novas funções para processar e para recuperar os artigos científicos, a partir de conceitos e de tecnologias da Web Semântica [123]. CISS-SW mapeia os termos de documentos em triplas RDF [124] e consultas em SPARQL [125], usando a história clínica dos pacientes. Usando a consulta, CISS-SW faz a busca na base de armazenamento de triplas para recuperar o artigo relacionado com o registro clínico.

Depois de utilizar as classes de processamento textual (classes verdes na Figura 4.4), as triplas RDF são geradas e armazenadas em repositório de armazenamento de triplas na Camada de Armazenamento da Figura 4.3. No CISS-SW, o processamento de cálculo de similaridade da classe *Similarity* implementa um novo método, explorando consultas SPARQL construídas a partir do resultado do processamento de registros clínicos. Como o sistema CISS, o CISS-SW processa cada artigo gerando os n -grams e os pesos dos termos. Depois, o sistema interage com a classe *ConceptRecognition* e os pacotes para identificar os conceitos dos artigos científicos. Finalmente, CISS-SW cria triplas RDF compostas por URIs, o identificador do artigo (PMID) e o *Concept Unique Identifier* (CUI) recuperado para cada conceito identificado do artigo. As

triplas resultantes são armazenadas no *Virtuoso Triple Store* [126]. Da mesma forma, o processamento dos registros clínicos de pacientes gera n -grams e distância euclidiana entre consultas SPARQLs de registros médicos e artigos para recuperar artigos científicos, que são semelhantes a um prontuário clínico.

O HSSF suportou perfeitamente o desenvolvimento do CISS-SW, apenas algumas classes como a *ConceptRecognition* e a *Similarity* foram modificadas para incluir novos métodos para criar e armazenar triplas RDF, e fazer buscas a partir de consultas SPARQLs.

Os resultados iniciais do CISS-SW foram publicados e apresentados em um poster no *MedInfo 2015* [123]. Os resultados comparativos CISS, CISS+ e CISS-SW estão em [13]. Esses resultados indicam a complementariedade de resultados devido ao uso de diferentes mecanismos de manipulação de informação, apesar da existência de algumas sobreposições.

4.4 Question Answering Surveillance Framework (QASF)

O QASF tem uma arquitetura do tipo *Question Answering*⁸ com o objetivo auxiliar profissionais da área da saúde respondendo questões do domínio de doenças crônicas. O QASF é composto pelos três módulos tradicionais de QA: (i) processamento da questão, (ii) processamento da resposta e (iii) processamento do documento; esse último módulo foi instanciado do HSSF. A Figura 4.5 apresenta a arquitetura do QASF.

Figura 4.5: Arquitetura geral de um sistema de *Question Answering* junto aos submódulos do QASF [15].

O módulo de *Processamento da Questão* extrai informações da questão formulada pelo usuário em linguagem natural. Essas informações são fundamentais para a seleção das respostas candidatas, e para converter a questão em uma consulta em linguagem computacional. A

⁸Os sistemas de informação do tipo Perguntas e Respostas ou *Question Answering* (QA) têm o objetivo de suportar sistemas que forneçam informações diretas e precisas sobre uma pergunta proposta pelo usuário. Para realizar esta tarefa, os sistemas de QA utilizam técnicas computacionais, principalmente, das áreas de Extração de Informação, Mineração de Texto e Recuperação de Informação [127].

consulta é utilizada, posteriormente, no módulo de *Processamento dos Documentos*. A etapa de “Processamento da Questão” é dividida em: *Análise da Questão* e *Elaboração da Consulta*.

Na literatura, foram encontradas algumas abordagens para “Análise da Questão”: correspondência de padrões [128], Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) e análise morfosintática [129, 130]. O QASF é suportado por abordagens de classificação de Aprendizado de Máquina, especificamente por SVM e Naive Bayes, para construção do submódulo de *Análise da Questão*. Os classificadores foram construídos e validados utilizando um conjunto de perguntas subdividido de acordo com o tema da pergunta (classes ou *tags*). No submódulo de *Análise da Questão*, uma coleção de perguntas dividida em temas (classes ou *tags*) conhecidos é utilizada para que o algoritmo de análise de questão classifique uma pergunta não analisada. O algoritmo deve ser capaz de acertar o tema (classe ou *tag*) da pergunta. No submódulo de *Análise de Questão*, uma coleção de perguntas dividida em temas (classes ou *tags*) conhecidos é utilizada para que o algoritmo de análise de questão classifique uma pergunta não analisada. O algoritmo deve ser capaz de acertar o tema (classe ou *tag*) da pergunta.

Os resultados obtidos com os classificadores foram comparados entre si. O classificador SVM apresentou melhor desempenho com valores acima de 0.50 para as medidas de Precisão, Revocação, F-Measure e Acurácia. Especificamente, no cenário cujas as classes que continham menos de cem questões foram descartadas, os valores da precisão e acurácia alcançaram valores próximos a 0.6. Desse modo foi possível perceber a importância do balanceamento das classes na tarefa de classificação da questão. Esta avaliação indicou performance moderada do QASF. Atualmente, os demais módulos estão em avaliação.

No submódulo de *Elaboração da Consulta*, um conjunto de palavras-chave é extraído, o qual será a base da consulta computacional, de uma pergunta inserida pelo usuário em linguagem natural. A representação da consulta em um vetor de palavras é transmitida para o módulo de processamento do documento, especificamente, para o submódulo de *Recuperação de Documentos*. No QASF, a construção do vetor de palavras é feita desconsiderando as *question words*, as *stopwords* e as pontuações. As *question words* têm grande importância na fase de “Processamento da Questão” para definir o tipo de pergunta inserida pelo usuário, no QASF, elas são retiradas da consulta, mas armazenadas em um arquivo separado. Como resultado do processamento da questão é construído um vetor com as palavras-chave da pergunta.

O módulo de *Processamento dos Documentos* engloba os submódulos de *Recuperação dos Documentos Candidatos* e de *Extração das Respostas Candidatas*. A tarefa de “Recuperação de Documentos” objetiva a recuperação de documentos que contenham possíveis respostas para a pergunta elaborada pelo usuário. Como entrada de dados desta etapa, a consulta criada em

um vetor de palavras na etapa de “Processamento da Questão”, a partir da pergunta feita pelo usuário. No QASF, o submódulo de *Recuperação de Documentos* foi projetado a partir do *framework* HSSF [30, 12, 11]. A ideia foi ampliar o leque das informações manipuladas para a busca de informações relevantes no contexto de um sistema de QA para obter mais precisão das respostas retornadas ao usuário, uma vez que o HSSF compreende os processos de busca da informação em artigos científicos e prontuários médico do paciente nos domínios de epigenética e doenças crônicas.

O submódulo de *Extração das Respostas Candidatas* é um dos mais complexos, pois o usuário pode fazer perguntas em diferentes níveis de complexidade. Na literatura, vários autores propõem diferentes abordagens como, por exemplo, utilizar as *tags* encontradas no submódulo de “Análise da Questão”, o uso de árvores de análise sintática ou de grafos de dependência [129] e a proximidade linear [127, 131]. No QASF, a extração de respostas candidatas utiliza os documentos retornados pelo submódulo de *Processamento do Documento*. Inicialmente, o conteúdo de cada artigo retornado é segmentado em parágrafos. Em seguida, uma comparação entre o vetor de palavras, que representa a pergunta inserida pelo usuário, e cada parágrafo segmentado dos artigos retornados pelo submódulo de *Recuperação de Documentos* é realizada. Como processamento, a quantidade de palavras contidas no vetor é verificada e se elas estão contidas no parágrafo candidato. Caso a quantidade seja igual ou maior a um limiar l , definido manualmente, o parágrafo é selecionado como parágrafo candidato.

O módulo de *Processamento da Resposta* é a última etapa do processo de *Question Answering*. Neste módulo, as respostas candidatas são selecionadas e ordenadas de acordo com o grau de similaridade entre a questão e as respostas candidatas. A etapa se divide entre os submódulos de *Seleção e Ordenação das Respostas* e *Validação das Respostas*. Como resultado deste módulo, as respostas para a pergunta inserida pelo usuário no início do processo de QA são retornadas. Os sistemas de QA não precisam, necessariamente, retornar uma única resposta, eles podem retornar várias respostas ordenadas de acordo com o grau de relevância. Atualmente, o QASF está retornando as três respostas mais semelhantes à pergunta inserida pelo usuário, contudo, esse valor pode ser reajustado.

O submódulo de *Ordenação e Seleção das Respostas* realiza a ordenação das respostas candidatas por meio de medidas de similaridade e apresenta ao usuário um número n de respostas mais semelhantes. Na ordenação das respostas candidatas, alguns critérios podem ser adotados: (i) a quantidade de palavras da questão do usuário, que, é reconhecida nas respostas candidatas na mesma ordem de posição, (ii) o número de palavras que separa as palavras-chave mais distantes entre si no parágrafo analisado ou (iii) o número de palavras incompatíveis com

Figura 4.6: Diagrama de classes do HSSF com o QASF.

as palavras-chave [127, 131]. Caso o sistema não encontre resposta, ele retorna na etapa de ordenação dos parágrafos. No QASF, o submódulo de *Ordenação e Seleção das Respostas* realiza o cálculo de similaridade usando a medida de cosseno. Nesse contexto, a pergunta feita pelo usuário e as respostas candidatas são convertidas para o espaço vetorial. Em seguida, a similaridade representada pelo cosseno entre a pergunta do usuário e cada resposta candidata é calculada. As respostas são ordenadas de acordo com o valor da similaridade e um número n de respostas mais similares são retornadas para o usuário.

O diagrama de classes do HSSF com as classes do QASF pode ser visualizado na Figura 4.6, estendendo o diagrama de classes do HSSF (com as classes no canto inferior direito). O sistema QASF foi aceito para publicação no *XVI Workshop de Informática Médica (WIM)* a ser realizado em julho de 2016 [15]. A modelagem de classes com experimentação deverá ser submetida em breve para um periódico internacional.

4.5 Sumarização de Artigos Científicos em Interações Gênicas (SARI)

Experimentos clássicos da área de genética revelaram que todas as células de um organismo possuem o mesmo conteúdo de DNA [132]. Apesar disso, as células de um organismo complexo se diferenciam e executam funções distintas e necessárias para a manutenção da vida do

organismo ao expressar genes diferentes, os quais são apropriados para cada situação, tecido, etc. Os genes são segmentos de DNA, que contêm as informações para codificar as proteínas e RNAs necessários para o funcionamento da célula. As células possuem diversos mecanismos para regular a expressão gênica, processo em que um gene sintetiza um produto, RNA ou proteína. Os padrões de expressão gênica se alteram de acordo com o estado fisiológico da célula, assim genes são ativados ou inativados nos processos de crescimento, divisão, respostas ao ambiente. Pode-se monitorar a expressão gênica utilizando técnicas de *microarray* de DNA. Os dados provenientes de um *microarray* representam o nível de atividade de milhares de genes simultaneamente em um ambiente bioquímico. A possibilidade de medir como os genes se comportam em um dado momento contribuiu para o entendimento de processos celulares, tratamento e diagnóstico de doenças e desenvolvimento de drogas [133]. Cada *microarray* possui grande quantidade de dados, por exemplo, um conjunto de dados de expressão gênica humana pode conter valores de expressão de até 39 mil genes [134][135]. Um dos principais objetivos da análise de *microarray* é agrupar genes com perfil de expressão gênica similares. Para a promoção de avanços científicos, a transformação dos dados gerados em informação e conhecimento é fundamental.

Clustering ou agrupamento é um tipo de aprendizado de máquina não-supervisionado utilizado na análise de *microarrays* de DNA. No *clustering*, dados são agrupados de acordo com similaridades, contudo métodos não-supervisionados exigem análises posteriores dos grupos gerados [136]. Na Internet, o enorme volume de dados e de literatura disponível dificulta a pesquisa de informações. Por exemplo, o desuso da nomenclatura gênica oficial é um dos problemas mais comuns em trabalhos científicos para a busca de informações de genes e seus produtos. Há casos de artigos com nomenclatura obsoleta, de autores que não especificam se fazem referência ao gene ou a proteína resultante, e etc [137]. Nesse cenário, a utilização de buscas avançadas, ferramentas de relacionamento automático de informações e sumários automáticos pode se tornar interessante. A sumarização automática busca extrair conteúdo de uma fonte de informação e apresentar somente o assunto mais importante. Considerando o grande volume de publicações científicas, a tarefa de identificar, selecionar e analisar textos de interesse tornou-se uma tarefa difícil. Assim, a utilização de sumários é um recurso interessante, uma vez que torna possível obter o conteúdo mais relevante de um texto, de forma condensada.

O método SARI (Sumarização Automática de Artigos Científicos para Representar o significado de Interações Gênicas) foi desenvolvido com objetivo de auxiliar na definição de significado a grupos de genes que interagem na maquinaria celular. Para alcançar esse objetivo o SARI foi proposto pela composição dos seguintes processos:

- Obtenção de dados de expressão gênica: a partir de bases de dados online (GEO (*Gene Expression Omnibus*) ou o *ArrayExpress*), mantidas pela própria comunidade científica, ou de experimentos realizadas em bancadas;
- Análise: a partir de métodos que permitam descrever fatos, detectar padrões e desenvolver explicações. Nesse processo técnicas de Processamento de Linguagem Natural e artefatos linguísticos como dicionários e ontologias são efetivamente explorados;
- Consulta a literatura científica: na busca de informações que expliquem (reforcem ou contradigam) os resultados da análise de dados do processo anterior. Ao estabelecer uma relação entre os dados de expressão com informações científicas externas, consegue-se agregar conhecimento ou fazer novas descobertas sobre os processos biológicos;
- Sumarização: para reduzir a quantidade de conteúdo das informações científicas relacionadas aos dados da análise de expressão gênica, sem que a informação principal do texto seja perdida;
- Apresentação sumarizada dos resultados científicos relacionados a expressões gênicas. Para visualização gráfica, optou-se por construir uma rede cujos os nós da rede são os genes e as arestas que ligam dois genes são os sumários dos artigos com as descrições das interações. Assim, obteve-se uma integração visual e textual, as quais auxiliaram na atribuição de significado aos agrupamentos.

Em relação aos materiais e métodos, diferentes abordagens de sumarização automática foram investigadas para verificar suas adaptações a artigos científicos que contêm as nomenclaturas de genes. Uma aplicação foi no auxílio ao processo de atribuição de significado aos agrupamentos gerados a partir de dados de expressão gênica. Quando a literatura científica indicava relacionamento entre genes de um agrupamento, pôde-se inferir que a classe não foi formada por aleatoriedade e que o algoritmo estava classificando, de acordo com estudos científicos previamente publicados. Nessa proposta, a sumarização foi guiada pela presença dos nomes dos genes nas sentenças.

Uma contribuição do SARI é o auxílio na validação e na atribuição de significado aos agrupamentos gerados a partir de dados de expressão gênica. Para isso considera-se a premissa que, quando a literatura científica indica relacionamento entre genes de uma classe, pode-se inferir que a classe não foi formada por aleatoriedade. Quando uma classe aponta um relacionamento entre genes, que nunca foram citados na literatura, isso pode indicar um novo foco de estudo ou um indicativo de problemas no algoritmo de *clustering*, ambos os casos devem ser analisados.

O método SARI foi apresentado no evento *X Meeting* em 2009 [138] e no Workshop de Informática Médica dentro do Congresso da Sociedade Brasileira de Computação em 2012 [139]. Esse trabalho deverá ter continuidade a partir do atual foco de investigação científica desta pesquisadora, o de Redes Complexas, apresentado no próximo capítulo.

4.6 Assistência de Gestantes para Vigilância de Fatores de Risco

Durante os últimos anos, a tecnologia móvel e os serviços computacionais tiveram grande avanço e ambos fazem parte da rotina diária das pessoas. O surgimento dos *smartphones* marcou o período de uma nova era, pois agora todas as pessoas podem ter um computador de bolso com capacidade de processamento semelhante, ou até mesmo superior, aos computadores do início da década. Esses celulares proporcionam inúmeras funcionalidades, e a principal delas é que se tornou possível experimentar, instalar e desinstalar aplicações, de um modo fácil e prático. Essa facilidade de acesso e uso de aplicações tem incentivado a desenvolvedores de todas as áreas do conhecimento a criar aplicações com diversas finalidades para o usuário final.

Na área da saúde e tecnologia, existe o conceito de mHealth como a geração, a agregação e a disseminação de informação de saúde via dispositivos móveis e com redes sem fio, como sistemas computacionais móveis que coletam, analisam os dados e fazem interação com os pacientes. As funcionalidades desse tipo de sistema promovem a comodidade e a satisfação do paciente, que não precisa sair de casa para, por exemplo, ver resultados e poder ter acompanhamento mais personalizado. Além disso, esses sistemas podem reduzir custos, já que sistemas mHealth podem, segundo pesquisas, diminuir o número de internações, tempo destas, e possíveis complicações [140].

No Brasil, há mais de 281 milhões de telefones celulares, sendo que existe cerca de 137 celulares a cada cem habitantes, e três *smartphones* para cada dez pessoas [141]. Enquanto o número de celulares ultrapassa a proporção de um para cada pessoa, o número de médicos é insuficiente: são apenas 1,95 médicos para cada mil habitantes no Brasil [142]. Esse cenário é motivador para o desenvolvimento de ferramentas de mHealth para vigilância da saúde.

Tendo em vista a problematização e o uso dos *smartphones* e o conceito de mHealth, esta pesquisadora e uma aluna de trabalho de conclusão do curso de Informática Biomédica estão desenvolvendo um protótipo de um sistema computacional, o Mommy, que visa monitorar e assistir gestantes durante o período de gestação prevenindo e/ou minimizando a exposição aos fatores de risco que sejam prejudiciais a saúde materna e fetal. Sabe-se que o período gestacio-

nal é um período que a gestante e o feto são extremamente vulneráveis, e muitos fatores podem colocar a gestação em risco [143, 144]. Segundo a literatura, esses fatores de risco podem ser divididos em categorias [145]: (1) condições de saúde existentes, (2) idade, (3) estilo de vida e (4) condições de gravidez. Esses riscos podem ser reduzidos quando tomadas as devidas medidas preventivas, no entanto, a taxa de mortalidade materna por causas relacionadas a gravidez e a taxa de mortalidade neonatal continuam altas [146, 147].

O trabalho busca apoiar a prevenção de complicações durante a gravidez, possíveis doenças na infância do neonato, ou até mesmo o óbito. O sistema também servirá para fornecer aos profissionais de saúde relatórios sobre a saúde da gestante, o qual o notificará de alguma anormalidade. Relatórios médicos possibilitarão ao profissional de saúde prestar um acompanhamento melhor e mais personalizado à gestante.

O sistema Mommy está sendo formado por: (i) um aplicativo Android que será instalado nos *smartphones* das gestantes, (ii) um sistema Web a ser utilizado pela equipe de saúde que faz o acompanhamento pré-natal da gestante, e (iii) um sistema central que será responsável pelo controle e pelo processamento das informações. Os *smartphones* servirão como interface de interação da gestante com o sistema e também serão utilizados para captar os dados do ambiente da gestante. Os dados capturados ou inseridos no sistema serão enviados para o servidor que possui um sistema central para processá-los e armazená-los. Após o processamento, o sistema enviará alertas e dicas para a gestante, e incluirá um resumo das informações no relatório para os profissionais de saúde. Esse relatório poderá ser visualizado por meio de uma interface Web. O projeto conta com as colaborações da Dr. Ligia Marçola, médica obstetra neonatologista da Santa Casa de Limeira e do Dr. Enrique Cifuentes da Harvard University.

A primeira versão do Mommy possuirá: o módulo de cadastro da gestante, o módulo de geração de relatórios para o profissional da saúde, o módulo de monitoramento da ingestão de medicamentos, as interface gráfica Web e do celular.

4.7 Considerações Finais

Reverendo os resultados, doze anos depois do início da carreira na USP e da incursão científica na área de Informática Biomédica, esta pesquisadora avalia positivamente sua trajetória. Uma oportunidade de trabalho que começou, apoiada pelos projetos FAPESP, com a oportunidade de orientador alunos mestres e doutores, porém na área de Aprendizado Eletrônico. Para suprir a falta de oportunidade de trabalhar em Informática Biomédica, esta docente se envolveu em projetos com alunos de graduação do curso (principalmente no projeto do Automatic-SL).

Em 2008, a oportunidade de participar de uma pós-graduação em Bioinformática fez a pesquisadora convergir os projetos que estavam em desenvolvimento com graduandos da Informática Biomédica para projetos da pós-graduação em Bioinformática. Assim sugeriram o CISS, o CISS+, o CISS-SW, o HSSF e o QASF na área de epigenética; o SARI em redes gênicas; e o FREDs (do Capítulo 3) em imagens citológicas. Todos os projetos, com alunos de graduação e de pós-graduação, geraram publicações em periódicos internacionais de destaque da área de Informática Biomédica como [11, 12, 9] e em congressos internacionais [30, 123]. Espera-se ainda resultados para dois artigos submetidos no contexto do HSSF [13, 14]. O QASF e o SARI foram publicados em congresso nacional, respectivamente em [15, 96]. Dando continuidade ao QASF, esta docente possui três alunos de graduação, dois da Informática Biomédica e um da matemática, apoiando o desenvolvimento a partir de bolsas de iniciação científica da Pró-Reitoria da USP pelo Programa Unificado de Bolsas. Ainda no ano de 2016, o projeto completo e experimentos do QASF serão apresentados em um artigo a ser submetido para periódico internacional. O SARI será retomado no contexto de Redes Complexas, apresentado a seguir.

Vale ressaltar que o trabalho publicado no periódico [11] foi resultado do projeto de graduação e de iniciação científica da aluna Juliana Pollettini no centro médico da Vila Lobato em Ribeirão Preto/SP. O trabalho de [123] é também oriundo de iniciação científica da aluna Julia Chaves, a qual está dando continuidade com o projeto Mommy com Havard. E finalmente, o projeto do FREDs começou na iniciação científica do aluno Hugo Pessotti [7] e foi finalizado no mestrado com publicação em periódico internacional [9]. Esses bons exemplos de continuidade e de produção desde a graduação motivam a pesquisadora a acreditar na importância do relacionamento das atividades de graduação e de pós-graduação na Universidade de São Paulo. Essa constatação faz a pesquisadora acreditar em um futuro ainda mais promissor, visto que em 2015 foram iniciadas as atividades da pós-graduação em Computação Aplicada no DCM-FFCLRP-USP, que administra o curso de Informática Biomédica.

5 *Manipulação e Relacionamento de Informações com Redes Complexas*

Redes (ou grafos) são conjuntos de nós (ou vértices) conectados por arestas (ou conexões, ou relações). Os grafos são dirigidos, quando as arestas entre cada dupla de nós apontam para algum sentido (são direcionadas), ou caso contrário, os grafos são não-dirigidos, pois possuem arestas sem apontamento de sentido. Exemplos tradicionais de rede são a Web, a Internet, as redes sociais, as redes biológicas e muitas outras. Recentemente, os estudos de redes tornaram-se viáveis para muitas pesquisas, principalmente devido ao destaque das redes reais com grandes números de nós e ao aumento da capacidade computacional. Esses estudos revelaram várias características que diferem as redes do mundo real das redes aleatórias, como eram consideradas as redes reais. Ao estudar as redes reais e suas propriedades, pesquisadores querem encontrar respostas para perguntas como: “a estrutura da rede pode afetar o tráfego na Internet ou o desempenho de um mecanismo de busca?”, “como é a dinâmica de sistemas sociais ou biológicos?”, “o conhecimento de uma rede social pode auxiliar na prevenção da propagação de uma epidemia ou de um vírus na Internet?”, e etc.

O termo Redes Complexas (RC) refere-se a um grafo que apresenta uma estrutura topográfica não trivial e as conexões entre seus nós não são puramente aleatória, nem puramente regulares [148]. RCs têm como principal característica o dinamismo de suas conexões entre nós. As pesquisas iniciais de redes em forma de grafos focavam nas pequenas redes e nas propriedades de vértices individuais ou arestas. As atuais pesquisas consideram redes reais com propriedades estatísticas em larga-escala e essas pesquisas indicam a não-aleatoriedade das topologias e evoluções das redes reais. A falta de padrão regular dessas redes reais em larga-escala caracteriza as redes complexas [148]. Essas redes apresentam algumas propriedades como coeficiente de aglomeração, conectividade ou grau, correlação de graus, ciclo, diâmetro, caminho mais curto, resistência e mistura de padrões. Além das redes aleatórias, outros tipos de redes são redes pequeno-mundo [149] e redes livres de escala [150]. O foco atual

de investigação científica desta pesquisadora é a modelagem das informações biomédicas, apresentada no Capítulo 4, em Redes Complexas. Conceitos gerais de RC e a modelagem inicial de uma Rede Complexa Heterogênea de Informação Biomédica (RCHIB) são apresentados neste capítulo.

5.1 Conceitos Gerais

Na literatura, principalmente na Matemática Discreta, as redes são representadas na forma de um grafo. Nesta seção, são apresentados alguns conceitos de suporte ao atual projeto em desenvolvimento por esta pesquisadora e seus alunos. As fórmulas matemáticas podem possuir diferentes representações na literatura.

Um grafo $G = (V, E)$ é uma estrutura de dados composta por um conjunto finito de nós, $V = V_1, V_2, \dots, V_m$, e um conjunto de arestas E , sendo $E = E_1, E_2, \dots, E_n$, ou $E \subset V \times V$. O tamanho (ou ordem) de um grafo G é definido como o número de nós em G , denotado como $n = |V|$.

Um grafo não-direcionado (ou não-dirigido) G é um par (V, E) , onde V é o conjunto de vértices e $V \neq \emptyset$, e consiste no par de vértices não-direcionado, isto é, (v_i, v_j) e (v_j, v_i) são a mesma aresta. Nesse tipo de grafo, dois vértices são ditos adjacentes se eles são pontos finais de uma mesma aresta. Por outro lado, um grafo direcionado ou (dígrafo) G é um par (V, E) , onde V é o conjunto finito de vértices e $V \neq \emptyset$ e E é uma relação binária em V , ou seja, as arestas $(v_i, v_j) \neq (v_j, v_i)$. Arestas têm uma direção associada. Em um grafo direcionado, um vértice V é adjacente a um vértice U se o par (U, V) é um arco, ou seja, se existe um arco que sai de U e entra em V . Também em grafos direcionados, existem *loops* que são uma aresta associada a um par de vértices (V_i, V_i) .

O grau d_i de um vértice V_i em um grafo não-direcionado é o número de arestas que incidem em V_i . O grau de um vértice em um grafo direcionado é o seu grau de entrada somado ao seu grau de saída. O número de arestas da rede, m , é o número de pares não-ordenados no conjunto E , ou seja $m = |E|$. A soma dos graus de todos os vértices de um grafo G é duas vezes o número de arestas de G , uma vez que cada aresta tem duas extremidades:

$$\sum_{i \in V} d_i = 2m \text{ ou } 2|E|.$$

O grau de um vértice é um valor entre 0 e $n - 1$, onde n é o número de nós ou vértices de G . Para obter o grau médio da rede, basta fazer a média aritmética do grau de todos os vértices:

$$\bar{d} = \frac{1}{n} \sum_{i \in V} d_i = \frac{2m}{n}.$$

Usando o número de vértices m e de arestas n , pode-se definir a densidade da rede ρ , que é a fração de arestas que a rede possui. A densidade máxima de uma rede ocorre se cada vértice tem o maior grau possível, $d_i = (n - 1)$, e consequentemente o total de arestas é $n(n - 1)/2$. Desse modo a densidade é:

$$\rho = \frac{m}{n(n-1)/2} = \frac{\bar{d}}{n-1}.$$

Um grafo $G = (V, E)$ é bipartido, quando seu conjunto de vértices V pode ser particionado em dois subconjuntos V_1 e V_2 , tais que toda aresta de G une um vértice de V_1 a outro de V_2 . Em grafo conexo, existe pelo menos um caminho entre todos os pares de vértices de G . Por outro lado um grafo desconexo é formado por dois ou mais grafos conexos, e cada subgrafo conexo é chamado de componente.

Algumas propriedades estatísticas comuns a Redes Complexas (RC) são:

- **Distribuição de Graus (DG):** é uma função de distribuição probabilística que indica a probabilidade de um vértice ter um determinado grau. DG é uma propriedade estrutural, de uma rede $G = (V, E)$ com p_k , sendo a fração de vértices com grau igual a k . Para quantificar essa distribuição, a função de distribuição complementar cumulativa do grau, P_k é a fração de vértices com grau maior ou igual a k . Obtém-se somando todos os graus menores do que k e fazendo o complemento:

$$P_k = 1 - \sum_{k'=k}^{\infty} p_{k'}.$$

Em um dígrafo, a fração de vértices tem simultaneamente graus de entrada j e saída k , p_{jk} . Grafos que representam redes reais são de livre escala com lei de potência para a distribuição de graus.

- **Comprimento Médio de um Caminho Mínimo (CM):** o comprimento de um caminho entre um par de vértices é igual ao número de arestas que formam o caminho que conecta os dois vértices. Existem diferentes caminhos que conectam um par de vértices. O caminho mais curto (ou mínimo, conhecido como distância geodésica) é o caminho com o menor número de arestas que conecta um vértice origem a um destino. A média do comprimento em uma rede é definida como a média do número de arestas dos caminhos mais curtos para todos os possíveis pares de nós conectados em uma rede. Nesse caso, o efeito mundo pequeno significa que para um certo valor fixado do grau médio, o valor da

média do comprimento de um caminho usa escala logarítmica, ou mais lenta, de acordo com o tamanho da rede. Assim, muitas redes apresentam alta conectividade formando pequenas quantidades de conexões em cada vértice. CM entre pares de vértices em um grafo não-direcionado é:

$$\frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij},$$

sendo d_{ij} é a distância geodésica do vértice i ao j . Essa definição apresenta problema nas redes com mais de um componente, uma vez que nessas redes há um subconjunto de vértices interconectado entre si, mas sem qualquer conexão com outro subconjunto da rede. Consequentemente, a distância média geodésica considera apenas os pares de nós em que há um caminho entre eles.

- **Coeficiente de Aglomeração (CA):** é uma medida de agrupamento intrínseco às redes, conhecido como Fenômeno de Transitividade. Esse fenômeno ocorre quando um vértice V_1 está conectado a um vértice V_2 , e o vértice V_2 está conectado a um vértice V_3 , aumentando as chances do vértice V_1 também estar conectado ao vértice V_3 . Assim, o CA indica a presença elevada de “triângulos” (vértices triplamente conectados — vértices com arestas não-direcionadas para o outros par de nós) na rede. CA é:

$$\frac{3T}{mV}.$$

O valor multiplicador três refere-se ao fato de que cada triângulo ser formado por três triplas, garantindo que o CA seja um valor entre zero e um, T representa o número de triângulos na rede e mV o número de vértices triplamente conectados. A média dos CAs em RCs reais tende a ser alta.

- **Comunidades:** uma comunidade é um grupo de vértice que tem alta densidade de arestas dentro do grupo (a comunidade) e uma baixa densidade de arestas para vértices de outros grupos (outras comunidades). Algumas redes têm comunidades ou estrutura de comunidade. Existem pesquisadores desenvolvendo técnicas para essa busca. Algumas vezes, as redes que têm estrutura de comunidade são referenciadas como redes de alto coeficiente de aglomeração, porém essas duas propriedades não são consideradas equivalentes. Além disso, tem sido observado que redes do mundo real apresentam uma emergente estrutura de comunidade.

Existem outras propriedades para RC como capacidade da rede em resistir a remoções de vértices sem perda de funcionalidade, correlação de graus e mistura de padrões. A mistura

de padrões é encontrada em redes cujos vértices são formados por diferentes tipos de objetos. Dado o domínio de seu trabalho, esta pesquisadora tem especial interesse em redes com mistura de padrões, as quais são denominadas redes heterogêneas e são apresentadas a seguir.

Em termos de modelagem de RC, o modelo de rede regular representa redes cujos todos os vértices possuem o mesmo grau. Porém na modelagem de dados do mundo real, poucas redes se comportam como rede regular. Exemplos de redes regulares são os modelos atômicos da área de Física. Os modelos mais influentes de RC são:

- **Rede Aleatória** de Erdős e Rényi (ER): é o modelo mais simples de uma RC [150]. Erdős e Rényi focaram em determinar as propriedades de grafos aleatórios utilizando argumentos probabilísticos. Considerando N como o número de vértices da rede, o número máximo de ligações que podem ser estabelecidas é:

$$N(N - 1)/2.$$

Nesse modelo, arestas não-direcionadas são adicionadas aleatoriamente entre os N vértices; e cada aresta independente tem uma probabilidade p de conectar dois vértices, assim o total de arestas que incidem em um determinado vértice é:

$$k = p(N - 1).$$

Esse modelo gera um grafo aleatório ER com N vértices e k arestas, definido como

$$G_{N,K}^{ER}.$$

Inicialmente com N vértices desconectados, o grafo ER forma-se conectando os vértices selecionados aleatoriamente até o número de arestas do grafo ser igual a k . Alguns pesquisadores definem o modelo de rede ER com uma distribuição de Poisson que é caracterizada por ter coeficientes de agrupamento baixo [151]. Erdős e Rényi concluíram que: uma rede tem aproximadamente a mesma quantidade de conexões e as mesmas chances de receber novas, e quanto mais complexa for a rede, maiores as chances dela ser ER [150].

- **Rede Pequeno Mundo** de Watts e Strogatz (WS): é caracterizada por redes com padrões altamente conectados, formando poucas conexões em cada vértice [149]. Às conexões entre os vértices mais próximos, os pesquisadores deram o nome de pequeno mundo. Desse modo, os comprimentos médios dos caminhos entre vértices não ultrapassam um número pequeno de vértices. O efeito pequeno mundo é observado nas redes em que a maioria dos vértices se conectam por um comprimento de caminho mínimo.

- **Rede Livre de Escala** de Barabási e Albert (BA): é representado por redes com ordem dinâmica de estruturação e com características específicas como conexão preferencial para novos nós (pouco *hubs* e muitos nós com poucas conexões) [150]. A representação matemática da rede segue uma função $f(x)$ que permanece inalterada com um fator multiplicativo sob um re-escalonamento dinâmico de variável independente x .

Os modelos apresentados não incorporam todas as quatro propriedades citadas anteriormente para RC, por exemplo, eles falham para incorporar a propriedade de estrutura de comunidade. O estudo de estrutura de comunidades em RC tem recebido atenção de diversos pesquisadores que buscam entender funcionalidades, estrutura e comportamento de comunidades. Além disso, os modelos de RC normalmente consideram redes com nós compostos por um único tipo de objeto. Vários modelos de RC estão sendo propostos para sanar as deficiências dos três modelos mais influentes, bem como novos algoritmos buscam a evolução da investigação de RC.

5.2 Redes Complexas para Investigação de Relacionamento de Informações

As aplicações de RC variam desde sua natureza (redes sociais, redes biológicas – neurais, proteínas, metabólicas e genéticas, redes de computadores, redes de informação entre outras) até sua finalidade. Alguns tipos de redes são denominados redes de informação, pois são caracterizadas pela interação entre informação e uma complexa estrutura de conexões subjacentes. Redes de informação têm como exemplos clássicos a rede das citações entre artigos científicos e a rede de ligações entre páginas Web, formando a WWW. Na análise de RC, pode-se utilizar artifícios de Recuperação de Informação (RI), de Processamento de Linguagem Natural (PLN), Mineração de Dados (MD), Mineração de Textos (MT) e outros.

Alguns pesquisadores subdividem as redes de informação em redes homogêneas, quando contêm apenas um tipo de informação em vértices e arestas e redes heterogêneas, quando incluem diferentes tipos de informação em vértices ou arestas [152, 153]. As redes homogêneas geralmente são extraídas de sistemas de interação real e a heterogeneidade de objetos ou arestas é simplesmente ignorada. Como o contexto de trabalho desta pesquisadora envolve informações textuais, imagens e outros objetos de dados, esta seção apresenta os atuais e os próximos trabalhos no contexto de redes heterogêneas de informação. Pretende-se extrair informações para construir RC de dados multimídia e/ou multimodais, dados em vários idiomas ou apenas dados simplesmente textuais e imagens de informações biomédicas. Esse último caso é o foco de tra-

balho atual desta pesquisadora e foi denominado Rede Complexa Heterogênea de Informações Biomédicas (RCHIB). No caso da RCHIB, tem-se a maioria das informações não-estruturadas e heterogêneas e algumas informações estruturadas. Portanto, a construção de RC de qualidade torna-se um desafio. Esse desafio é ainda maior quando se pretende apoiar o relacionamento dessas informações em nós ou arestas usando artefatos linguísticos como ontologias, dicionários e Linked Data¹. Redes heterogêneas são apresentada a seguir e a RCHIB na Seção 5.3.

5.2.1 Redes Complexas Heterogêneas

Redes heterogêneas de informação fornecem um paradigma atual e desafiador para manipulação de informações e de dados em rede. Os tradicionais estudos sobre RC costumam se concentrar nas estruturas, funções e características dessas redes. Porém, uma RC heterogênea de informação também apresenta conteúdos de nós com inestimável valor. Esses conteúdos normalmente advêm de diferentes fontes de informação, que fundem informações por meio de múltiplas “plataformas” com diferentes objetivos. Além dos diferentes tipos de informações relacionadas por suas arestas, uma rede heterogênea possui como importante característica o valor semântico de suas arestas. Assim, as redes heterogêneas podem aproveitar a complementaridade das diferentes fontes e tipos de informação para melhorar a semântica da recuperação de informação. As redes de informação heterogêneas também podem efetivamente fundir informações, constituindo uma nova visão do conteúdo. Porém, a combinação de tipos de informações que estão em diferentes níveis de apresentação ou mesmo de semântica, tais como textos e imagens, é um exemplo do problema denominado *gap* semântico (ou brecha semântica). Esse *gap* semântico pode ser contornado com a complementariedade semântica das arestas em redes de informação heterogêneas. Para alguns pesquisadores, a complexidade das redes heterogêneas está no fato de trabalhar a ambiguidade dessas redes [148].

Youssef identifica dois tipos de heterogeneidade, a heterogeneidade dos atributos do nó e a heterogeneidade dos padrões de conexão do nó [154]. A heterogeneidade dos atributos do nó reflete as diferentes propriedades ou características de diferentes nós da rede. A heterogeneidade dos padrões de conexão reflete a diferença de requisitos de cada nó para fazer uma conexão com outro nó.

Considerando que a estrutura complexa e a semântica rica das redes de informação heterogêneas são as mais importantes características de redes heterogêneas, deve-se focar esforços e pesquisas [155]:

---

¹<http://linkeddata.org/>

- na estrutura da rede, uma vez que os objetos podem ser organizados de diferentes formas. Por exemplo, grafos bipartidos são usados para organizar diferentes tipos de objetos e os relacionamentos entre eles [156, 157, 158]. Redes em camadas organizam nós de maneira distinta [159, 160].
- no significado semântico dos objetos e ligações. Por exemplo, o uso de Meta-path pode efetivamente capturar a semântica entre objetos [161].

Em termos de modelagem de redes heterogêneas, há uma necessidade importante para a elaboração de um modelo matemático que facilite a realização de análise matemática em RCs desse tipo [162]. Esses modelos matemáticos podem ser utilizados para observar e/ou prever a forma como a rede se comporta sob diferentes cenários. Os modelos matemáticos também podem ser utilizados, quando os conjuntos de dados reais são impossíveis ou dispendiosos para análise da rede. Um bom modelo matemático deve imitar com sucesso propriedades estatísticas da rede modelada.

Youssef propôs dois modelos matemáticos para gerar RCs heterogêneas [154]:

- o *Integrated Attribute Similarity Model* (IASM) que incorpora a heterogeneidade dos nós, integrando atributo de semelhança com a medida de popularidade estrutural dentro da função de criação de ligação preferencial. A similaridade de atributo é utilizada para avaliar a semelhança ou compatibilidade entre os atributos de ambos os nós a serem conectados. Em contraste, a popularidade estrutural mede a popularidade dos nós mais velhos com base em suas conexões atuais. Popularidade estrutural pode ser baseada no número de primeiros graus de conexões do nó (grau de centralidade) ou nas conexões de graus mais elevados (centralidade por autovetores).
- o *Settling Node Adaptive Model* (SNAM) que usa um algoritmo de conexão com base em padrões de conexão para nós e não usa o algoritmo de criação de ligação preferencial como a maioria dos algoritmos que analisam propriedades de atributos para definição de relacionamentos.

Os modelos de Youssef e outros modelos para RCs heterogêneas estão sendo profundamente investigados pela pesquisadora e seu grupo.

5.3 Redes Complexas Heterogêneas para Informações Biomédicas

Os modelos de RC da Seção 5.1 buscavam definir modelos matemáticos para preservar propriedades estatísticas de redes do mundo real. No entanto, modelos mais recentes concentram-se na modelagem da criação e evolução da rede [154, 163]. As abordagens de modelagem da evolução da rede investigam como certas propriedades estatísticas surgem em redes do mundo real. Alguns pesquisadores incluem nessas abordagens modernas apenas a heterogeneidade dos nós; outros consideram importante incluir a heterogeneidade de nós e das arestas [154]. Esta pesquisadora considera, que além de incluir heterogeneidade dos nós, as pesquisas de RC heterogêneas modernas se fortalecerão com o desenvolvimento de padrões de relacionamentos de nós, usando a heterogeneidade das ligações, para criar condições para a geração de conhecimento a partir de ligações seletivas. Em outras palavras, nós tendem a conectar-se com os nós que são semelhantes a eles, em algum aspecto. Por exemplo, o assunto desempenha papel importante para conexão entre páginas Web. Ao tratar da semântica do conteúdo do nó, diferentes ligações podem ser criadas entre documentos. Acredita-se que essa investigação de ligações por assunto possa gerar comunidades ou grupos de vértices que são semelhantes de alguma forma e tendem a ter conexões densas entre si em dado assunto e conexões menos densas com nós pertencentes a diferentes comunidades. Nesse caso de ligação seletiva, trata-se a heterogeneidade do padrão de ligação do nó como a diferença nos requisitos de cada nó para fazer uma conexão.

Na área da saúde e da medicina, diferentes fontes de informação publicam informações de diferentes tipos com diferentes significados semânticos na Web via bibliotecas digitais, em sites de entidades públicas e privadas de saúde e medicina, em páginas pessoais, em redes sociais, em jornais e revistas, em sites de comunidades científicas entre outras. Em repositórios privados de entidades de saúde, também existem informações como registros médicos que poderiam se beneficiar ao serem relacionadas com algumas fontes de informação confiáveis na Web. A rede de informação formada pelas fontes de informação citadas apoiadas por entidades de informação de apoio como artefatos linguísticos (UMLS, ontologia, *Open Linked Data* etc) pode formar uma rede complexa heterogênea de informação devido à sua estrutura complexa e a rica semântica. Essa rede de informação biomédica seria composta por nós (ou entidades de informação) formados por diferentes tipos de objetos que podem gerar relacionamentos a partir de diferentes padrões de ligação, guiando-se por algum aspecto. Esse tipo de rede pode revelar importantes contribuições e conhecimento como, por exemplo, o auxílio que informações científicas podem prestar à comunidade da área de saúde na prevenção de doenças, a definição de comunidades de risco e outras. Portanto, a investigação de Redes Complexas Heterogêneas

de Informações Biomédica (RCHIB) é o foco atual e futuro de investigação desta pesquisadora. Os trabalhos prévios de manipulação, de relacionamento e de classificação de informação biomédica na prevenção a doenças estão suportando as atuais e as novas pesquisas do grupo.

5.3.1 Materiais e Métodos

A heterogeneidade das características de um nó poderá ser apresentada em um modelo teórico de grafos por meio da incorporação de atributos de um nó como um dos elementos de um grafo de rede. Assim, o grafo G poderá ser definido como um conjunto de três elementos, $G = V, E, A$, cujo V é o conjunto de nós ou vértices na rede, E é o conjunto de arestas ou relacionamentos, e A é o conjunto de vetores de atributos atribuídos a cada nó da rede. O comprimento de cada vetor em A é geralmente mais do que a unidade e não restrito à unidade para representação de RCs heterogêneas. Por outro lado, a heterogeneidade do padrão de ligações pode ser modelada como uma função de mapeamento $f : E \rightarrow R$, sendo que cada ligação $e \in E$ pertence a um tipo particular de relacionamento no conjunto de relações R .

Nós, ou entidades, da RCHIB tem diferentes propriedades e atributos, que influenciam os padrões de conexões dentro da rede. O conjunto de atributos de cada nó é formado pelos conteúdos-chave da entidade e suas propriedades como, a data de sua criação, a atratividade, o grau, a idade do nó, a capacidade do nó e outros metadados. Para esta pesquisadora, inicialmente a RCHIB é representada pelas seguintes entidades (ou nós) com os seguintes exemplos de atributos:

- Artigos científicos: ID, palavras-chaves, título, resumo, citações, referências, popularidade, grau, idade, autores, conferências e revistas;
- Imagens: ID, legenda, nome, informações próximas a citação da figura em texto, idade, popularidade, grau e outros;
- Laudos médicos: ID, atributos do texto, atributos das imagens, responsável e outros);
- Registros médicos de pacientes (ID, campos de informação e outros;
- Registros de informação em redes sociais: ID, campo de informação e outros;
- Ontologias: ID, conceito, significado, estrutura e outros;
- Dicionários: ID, conceito, significado e outros;
- *Linked Open Data Repository*, etc.

O três últimos vértices estão sendo considerados para suporte das tarefas de mineração da informação da rede. A Figura 5.1 apresenta a modelagem inicial de entidades e relacionamentos da Rede Complexa Heterogênea de Informação Biomédica - RCHIB, em criação por esta pesquisadora e seus alunos. A figura ilustra a rede com seus vértices, arestas e camadas (conjuntos) de vértices, os quais foram agrupados por contexto em camadas. Na rede, cada nó representa um tipo de entidade citada anteriormente. O domínio principal no contexto da RCHIB é o de informações biomédicas que fica na camada superior da figura. As outras duas camadas sustentam a camada superior e permitirão a busca filtrada de informações por camada.

A Figura 5.1 é composta por quatro grupos de vértices. Os vértices de cor verde são agrupados como documentos biomédicos (*Prontuários Eletrônicos, Pacientes, Laudos, Doenças e Imagens Médicas*) e são relacionados entre si por arestas de cor azul claro. Os vértices de cor lilás são vértices relacionados a conceitos que envolvem a publicação de Artigos científicos escritos por Autores para serem publicados em Revistas e Conferencias. As arestas de cor azul escuro relacionam esses conceitos do contexto de publicação científica. Finalmente, os vértices na cor azul representam aplicativos de redes sociais e científicas na Web (WebNars, Researchgate, Gplus, Flickr, Twitter, Facebook, Youtube, Linkdin). Esses aplicativos estão agrupados e relacionados entre si por arestas vermelhas. Finalmente, o vértice amarelo agrupa artefatos linguísticos a serem explorados pelos vértices e arestas da rede para suportar os relacionamentos.

Nas camadas, cada nó da rede tem um atributo-extra para indicar a camada que determinado nó pertence. Essa abordagem pretende facilitar a busca por conteúdo em contextos específicos (grupos de documentos). Portanto, uma busca específica em um grupo de documentos, elimina a necessidade de busca em toda a rede. No caso das redes heterogêneas com camadas, as linhas tracejadas representam relacionamentos entre as camadas.

Em termos de arestas, a Figura 5.1 apresenta as arestas tracejadas na cor azul claro, as quais relacionam imagens de documentos biomédicos e do aplicativo Flickr. As arestas tracejadas em verde claro relacionam informação sobre doenças a partir de ontologias ou outros artefatos linguísticos, com imagens e suas informações encontradas em artigos científicos de revistas e de conferências. As arestas tracejadas em rosa relacionam textualmente artigos científicos com prontuários eletrônicos e aplicativos como Twitter Facebook e Gplus. No caso de prontuários eletrônicos e artigos científicos, as arestas tracejadas na cor rosa associam artigos com prontuários eletrônicos, que são usados como base de informação para publicações de artigos científicos, normalmente para artigos do tipo discussão de casos. As arestas em vermelho relacionam a divulgação de forma escrita de artigos por meio do Twitter, Facebook e Gplus.

Figura 5.1: Entidades e Relacionamentos iniciais da Rede Complexa Heterogênea de Informação Biomédica.

As arestas em verde escuro relacionam as revistas científicas que divulgam suas imagens (de base de dados/ figuras estatísticas) no Flickr. As arestas em roxo relacionam os autores que divulgam seus trabalhos no Gplus, Twitter, e no Facebook. Já as arestas tracejadas em marrom relacionam autores que utilizam o ResearchGate para divulgar suas pesquisas. Finalmente, as arestas tracejadas em azul escuro associam vídeo de autores em conferências por meio de redes sociais para redes sociais.

Portanto, a Figura 5.1 ilustra os relacionamentos de uma rede heterogênea de informação composta por documentos biomédicos, artigos científicos e informações relacionadas a saúde publicadas em redes sociais. Essa rede é uma abstração de relacionamentos de uma rede real entre os documentos citados. A rede da figura possui arestas sem peso e não direcionadas para facilitar o entendimento da modelagem.

Na RCHIB, as redes sociais são os vértices mais interconectados, uma vez que elas têm como foco o compartilhamento de informação. A Figura 5.1 não apresenta todas as ligações dessas redes. Por exemplo, o ResearchGate é um rede da comunidade científica que compartilha artigos/documentos, além de permitir a criação e a resposta de perguntas e a busca de colaboradores. Assim, cada uma dessas atividades em uma rede completa exibiria ligações direcionadas para autores e para artigos. Esses tipos de ligações não estão presentes na Figura 5.1, porém elas estão sendo consideradas no projeto. Ainda em relação ao ResearchGate, cada artigo citado pode ser compartilhado em cada rede social, Facebook, Gplus e Twitter. Essas ligações estão representadas em cor vermelha na figura.

5.3.2 Metodologia

A modelagem inicial da rede complexa heterogênea de informações biomédicas inicialmente poderá utilizar o modelo BA, uma vez que o modelo BA considera o crescimento do tamanho da rede e faz conexões considerando aspectos de preferência, em vez de uniformidade ou aleatoriedade. Alguns modelos propostos para gerar RCs heterogêneas na literatura não são gerais, porque eles fazem suposições específicas sobre as propriedades da heterogeneidade da rede. Além disso, eles não preservam as quatro propriedades estatísticas de RC. Assim, a formulação de um modelo para a geração de redes complexas heterogêneas gerais com características que se assemelham tanto quanto possível as propriedades estatísticas comuns às redes do mundo real têm recebido a atenção da comunidade científica. Porém, essa é uma questão de pesquisa ainda em aberto.

Mecanismos computacionais de IR e de DM deverão ser desenvolvidos para análises e a criação da RCHIB. A maior parte dos atuais mecanismos na literatura é voltada para redes

homogêneas e não pode ser diretamente aplicada a mineração de redes de informação heterogêneas como a RCHIB, principalmente devido aos seguintes fatos: (i) ligações heterogêneas entre entidades de tipos diferentes podem ter significados semânticos diversos; e (ii) uma rede de informação heterogênea captura informações mais ricas que as redes homogêneas [161].

Atualmente, as entidades e os seus relacionamentos nos sistemas (Automatic-SL, CISS, CISS+, CISS-SW, QASF, FREDs e Mommy) de manipulação, relacionamento e classificação de informação biomédica na prevenção de doenças constituem o exemplo inicial de rede complexa heterogênea usado na RCHIB. Os sistemas prévios apresentam não só diferentes tipos de entidades (por exemplo, registros médicos, laudos, imagens e artigos científicos), mas também ilustram distintos tipos de relações entre os objetos, tais como visualização de informações, recomendação de artigo científico, atribuição de grau de vigilância, resposta a pergunta de usuários e recomendação de laudo a partir da análise de atributos de imagens. Essas relações foram basicamente suportadas por tarefas de mineração, do tipo classificação, e medidas de similaridade estatísticas. Imagina-se que os objetos e as relações nessas redes têm outras semânticas diferentes, que podem ser exploradas para revelar novas relações entre os objetos e gerar novas informações. Por exemplo, um caminho de pacientes relacionados ao mesmo artigo científico pode ser usado para encontrar outros pacientes semelhantes, de acordo com registros de visualização. Se artigos científicos forem recomendados seguindo esse relacionamento, ele irá recomendar os artigos que estão sendo vistos pelos usuários que têm os mesmos registros de visualização com o dado do utilizador (filtragem colaborativa). Da mesma forma, um relacionamento entre pacientes pode encontrar os pacientes semelhantes com problemas semelhantes (recomendação de membro). Assim, nota-se que é possível recomendar itens com base nos usuários similares gerados por diferentes caminhos para conectar os usuários e fazer detecção de comunidade. Além disso, pode-se perceber a existência de diferentes modelos de recomendação. Outras análises possíveis serão a previsão de ligação e a análise de influência. A mineração de ligações é uma área de intersecção de análise de ligações, hipertexto e mineração de Web, relacionando aprendizagem relacional e programação lógica indutiva, e mineração de grafos. Para criar essas novas arestas (relações) na RCHIB, pretende-se utilizar tarefas de mineração de informação como as seguintes:

- recomendação com uso de técnicas de recuperação de informação, estatística e aprendizagem de máquina para procurar semelhanças entre itens e principalmente atributos dos usuários. A filtragem colaborativa é uma das técnicas mais populares. Com a prevalência de mídia social, pesquisadores estudam informações captadas nessas mídias para filtragem de recomendações. Recentemente, alguns pesquisadores perceberam a importância da heterogeneidade informações para recomendações [164, 165, 44, 166].

- fusão de informação como processo de junção de informações de fontes heterogêneas com diferentes representações conceituais e contextuais. A disponibilidade de várias fontes de dados e informações motiva a fusão de fontes de informação distribuídas. Este tema é investigado por pesquisadores de mineração de dados, mas principalmente por pesquisadores de recuperação de informação. Para fundir a informação de múltiplas redes de informação heterogêneas um pré-requisito importante será alinhar as redes por meio de entidades de informações comuns, que podem ser pacientes, doenças ou artigos científicos. Ao fundir informações de diferentes redes heterogêneas de informação, esta pesquisadora imagina obter uma fonte de conhecimento mais abrangente e consistente sobre as entidades de informações comuns compartilhadas em diferentes redes, incluindo as suas estruturas, as propriedades e as atividades.
- medidas de similaridade para avaliar a similaridade de objetos. Essas medidas são a base para tarefas de busca na Web, recomendação de informação, classificação e categorização. As medidas de similaridade podem ser classificadas em medidas baseadas em atributos (similaridade de atributos de objetos como cosseno, coeficiente de Jaccard e distância Euclidiana) e medidas baseadas em relacionamento (similaridade de objetos de acordo com a estrutura representada normalmente em grafos ex. PageRank [35]). Existem autores propondo medidas de similaridade para vértices e relacionamentos em redes de informações heterogêneas [167, 168].
- *clustering* como um processo de particionamento de um conjunto de objetos em um conjunto de agrupamentos, de modo que os objetos em um agrupamento sejam semelhantes entre si por algum aspecto, mas diferentes de objetos em outros agrupamentos. O *clustering* convencional baseia-se no agrupamento de objetos devido as suas características. Recentemente, o agrupamento com base em dados de RC (por exemplo, detecção de comunidades) tem sido estudado.
- classificação como uma tarefa de análise para prever a classe que um objeto pertence. Aprendizado de máquina tradicional tem-se centrado sobre a classificação de objetos. Métodos de aprendizado de máquina ou de outras áreas podem estender os tradicionais métodos de classificação para considerar as correlações entre os objetos. Alguns pesquisadores estão trabalhando nessa linha [169, 170, 171].
- predição de ligações como um problema fundamental na mineração de ligações na tentativa de estimar a probabilidade da existência de uma relação entre dois nós, com base em ligações observadas e nos atributos de nós. Previsão de ligação pode ser um problema de classificação binária simples: para quaisquer dois objetos potencialmente vinculados,

prever se existe a ligação (1) ou não (0). Porém uma abordagem mais abrangente pode fazer a predição baseado em propriedades estruturais da rede ou em atributos dos nós.

- ordenação (ranking) como uma tarefa para análise de rede, uma vez que avalia a “importância” ou a popularidade do objeto com base em algumas funções de classificação. Exemplos clássicos são os algoritmos PageRank [35] e o HITS [36].

A adaptação ou a criação das tarefas das mineração citadas é um foco de investigação e contribuição da RCHIB. Alguns outros desafios que a RCHIB deve acolher são:

- propagação de informações por meio de ligações heterogêneas, uma vez que elas devem ser usadas para esse fim. No entanto, muitas dúvidas deverão ser sanadas como por exemplo, “como propagar informações entre os tipos heterogêneos de nós e ligações?”, “Usar ranking, classificação, categorização?”, e “Como usar os rótulos de classes, em nós e ligações heterogêneas?”. Não importa as respostas para essas dúvidas, mas uma regra de ouro deverá ser seguida: os objetos nas redes são interdependentes e o conhecimento poderá ser extraído usando a informação global.
- pesquisa e mineração, explorando meta-estruturas da rede [44]. Diferentemente de redes homogêneas, as redes heterogêneas são semi-estruturadas e tipadas como a RCHIB. Essa semi-estruturação pode orientar a mineração e ajudar a análise e compreensão do significado semântico dos objetos e das relações da rede.
- exploração guiada pelo usuário na rede. Em uma rede de informação heterogênea, existem numerosas relações semânticas para vários tipos de objetos. Assim é interessante selecionar automaticamente combinações de relações diretas com pesos adequados para uma determinada tarefa de busca ou de mineração e apoiar essa navegação (exploração) nas orientações ou comentários do usuário.

Para a área de saúde e de medicina, existem poucas propostas redes de informação heterogêneas como a RCHIB. Elas usualmente modelam o sistema de saúde contendo um conjunto de tipos de objetos, como médico, paciente, doença, tratamento, e um conjunto de tipos de relações, como utilizada para-entre para tratamentos/doenças e pacientes, ou visita de pacientes a médicos. Esse nível de detalhamento de entidades (nós) poderá ocorrer no futuro, de acordo com a necessidade da RCHIB.

5.4 Considerações Finais

Após quinze anos de estudo e de contribuições teóricos-práticos a cerca de relacionamento e de manipulação de informação, esta pesquisadora possui a chance de representar formalmente esse domínio de investigação usando conceitos da área de Redes Complexas Heterogêneas em Informática Biomédica. Acredita-se que a exploração do tema nesse domínio deverá permitir modelagens precisas, análises mais eficientes e resultados mais confiáveis. A associação de informações a partir de diferentes fontes poderá contribuir com a geração de conhecimentos que potencialmente deverão ser utilizados para a promoção da saúde de pacientes.

6 Conclusão

Devido ao fato de ter iniciado sua carreira científica em um grupo de pesquisa com foco de contribuição nas áreas de hipermídia, de multimídia, de interação usuário-computador e Web, desde suas remotas investigações científicas nessas áreas, esta pesquisadora foi motivada por Bush com seu Memex de relacionamento de informação, por Tim Bernes-Lee e sua proposta da Web, por Engelbart, Nielsen e Norman voltados para o uso da tecnologia para resolução de maneira simples de problemas de usuários sem habilidades específicas. Dessa mistura de motivações e áreas, as pesquisas teórico-práticas para a definição de mecanismos computacionais para o relacionamento de informações começou de maneira quase fortuita e por curiosidade pessoal em um dos estágios desta pesquisadora na Georgia Tech em Atlanta nos EUA em 1999. As ferramentas de aprendizado eletrônico dos grupos de pesquisa daquela universidade, a CoWeb e o iClass, possuíam informações complementares, porém os alunos não podiam beneficiar-se do conteúdo de ambas, ao mesmo tempo e de modo complementar. Dessa semente, surgiu uma proposta de doutorado em 2000 com apoio da FAPESP para investigar o relacionamento e a extensão de informações no contexto de um projeto de colaboração internacional com NSF (*National Science Foundation*) nos EUA. O projeto foi desenvolvido no período de quatro anos com publicações anuais em eventos nacionais e internacionais relacionados as áreas de hipermídia, Web e engenharia de documentos. Nesses quatro anos, o grupo de pesquisa no Brasil obteve infraestruturas para criação de ambientes ubíquos para os quais os serviço de ligações também foram integrados. Essa colaboração e resultados impulsionaram o início da carreira profissional da pesquisadora dando suporte aos Projetos Jovem Pesquisador e TIDIA-Ae, apoiados pela FAPESP, nos quais esta pesquisadora foi coordenadora a partir de 2004.

Em 2004, esta pesquisadora, recém-doutora e recém-contratada para o curso de Informática Biomédica na FFCLRP-USP, teve a oportunidade de iniciar um caminho multidisciplinar nas áreas de computação e de saúde para suas incursões científicas. Esse caminho com termos e temas de rica semântica era completamente desconhecido para a pesquisadora e esse desafio

inédito gerou novos aprendizados e ajustes das investigações científicas prévias para contribuição em saúde, a qual é a principal área de atuação do campus da USP na cidade de Ribeirão Preto. No contexto da saúde, a pesquisadora focou contribuições no paciente, principalmente, na prevenção de doenças. A ingreme curva de aprendizado e a continua extensão do tema de relacionamento de informações trouxeram oportunidades da criação de parcerias nacionais e internacionais e do desenvolvimento de trabalhos com resultados inéditos publicados em excelentes eventos e periódicos na computação e na informática biomédica. Em 2007, a possibilidade de participação em um programa de pós-graduação, o de Bioinformática, possibilitou a orientação de alunos de mestrado, que desenvolveram, por exemplo, o tema de vigilância de doenças suportada por pesquisas científicas em busca de alternativas de prevenção de doenças crônicas. Porém, o fato de participar de um programa em Bioinformática abriu novamente um campo de pesquisa a ser explorado por esta pesquisadora, que explorou epigenética e imagens citológicas. Em 2015, inicia-se o programa de pós-graduação em Computação Aplicada do Departamento de Computação e Matemática da FFCLRP-USP. Uma das linhas do programa é a de sistemas complexos, os quais norteiam os atuais e futuros anseios de investigação científica desta pesquisadora.

Além dos resultados publicados, os projetos, os benefícios de projetos da pesquisadora e as bolsas de seus alunos trouxeram recursos e possibilidades de grande interesse para o departamento, para a unidade, para a universidade e para a comunidade. Por exemplo, dois alunos de mestrado desta pesquisadora criaram a empresa Kidopi¹, a qual tem sido nacional e internacionalmente premiada por seus projetos na área da saúde. Outro exemplo de contribuição na comunidade foi o projeto com o centro médico de Vila Lobato, o qual gerou resultados expressivos de classificação de pacientes em grupos de atendimento, diferenciando cuidados rotinas de cuidados emergenciais. Os recursos possibilitaram as colaborações internacionais com profissionais da Georgia Tech, da University of Wisconsin, da Universidade de Coruña, da Auckland University e atualmente da Harvard University no projeto do Mommy. Em termos de colaborações nacionais, os projetos realizados contaram com participações de pesquisadores e alunos do ICMC-USP, da FMRP-USP, do Instituto de Biociências da USP, do ICR do HC-FMUSP, da Unifesp, da UFG, da UFU, do Inpa pelo projeto Adapta e de outras. Em relação a colaborações entre professores do próprio departamento, esta pesquisadora destaca que dos nove atuais professores do grupo de computação, seus trabalhos tiveram a colaboração de sete pesquisadores do grupo. Dessas colaborações, o maior destaque se dá à colaboração com o Prof. Dr. José Augusto Baranauskas pelo permanente e recíproco trabalho em conjunto dos grupos orientados pelos dois professores. O pesquisador apoia fortemente o desenvol-

---

¹<http://kidopi.com.br/>

vimento e o uso de mecanismos de Inteligência Artificial e Aprendizado de Máquina, além dos métodos envolvidos na experimentação das pesquisas. Por outro lado, esta pesquisadora apoia o grupo do Prof. Baranauskas na modelagem e na preparação dos artefatos de software. Essa colaboração mútua têm gerado excelentes resultados publicados ao longo dos últimos sete anos [172, 173, 13, 174, 12, 68, 11, 175, 176, 177, 99].

As pesquisas e os resultados apresentados neste texto envolvem temas atuais, multidisciplinares e de grande interesse na comunidade científica. Inicialmente, o foco foi a resolução de problemas na Internet, especificamente na associação de documentos Web (Capítulo 2); esse foco de pesquisa foi expandido para ambientes ubíquos com documentos multimodais e de imagens (Capítulo 3). A vinda para a USP de Ribeirão Preto convergiu os objetivos para a busca da efetiva transformação dos resultados de pesquisas biomédicas em conhecimento possível de ser utilizado para beneficiar a saúde pública (Capítulo 4). Recentemente, a criação do programa de pós-graduação em Computação Aplicada a Biociências e a Sistemas Complexos motivou o projeto de criação de uma rede complexa heterogênea para informações biomédicas (Capítulo 5). Como poucas tarefas tradicionais de mineração de dados e de informação podem ser aproveitadas em redes heterogêneas de informação, este campo de pesquisa é considerado jovem e promissor, fornecendo motivações para trabalhos futuros da pesquisadora e seu grupo de alunos e de colaboradores.

6.1 Dificuldades Encontradas

A vida acadêmica desta docente possui alguns aspectos laboriosos, os quais demandaram reflexões, mudanças e novos desafios. Alguns desses aspectos são: (i) o foco inicial para publicação de resultados de pesquisas era os congressos internacionais, principalmente no período de doutoramento; (ii) o ingresso em um Departamento de Física para trabalhar no curso Interdisciplinar de Informática Biomédica; (iii) a falta de um curso de pós-graduação na área no departamento ao qual a pesquisadora foi inicialmente incorporada; (iv) a participação em um programa de pós-graduação em uma área totalmente desconhecida pela pesquisadora, a de Bioinformática e (v) a criação recente de um programa de pós-graduação no departamento, apenas com a possibilidade de orientação de alunos de mestrado.

Referências

- [1] A. A. Macedo, *Especificação, instanciação e experimentação de um arcabouço para criação automática de ligações hipertexto entre informações homogêneas*. PhD thesis, Universidade de São Paulo, 2004.
- [2] R. d. F. B. Neto, J. A. C. Guerrero, P. Schor, A. S. Lopes, M. B. Dutra, and A. A. Macedo, “Capturing and relating multilingual clinical cases,” *Journal of Universal Computer Science (Print)*, vol. 20, pp. 1154–1173, 2014.
- [3] A. A. Macedo, L. A. Baldochi, J. A. Camacho-Guerrero, R. G. Cattelan, , and M. G. C. Pimentel, “Automatically linking live experiences captured through a ubiquitous infrastructure,” *Multimedia Tools and Applications (MTAP) Journal*, pp. 93–115, 2008.
- [4] R. Bulcão-Neto, A. Macedo, L. Wichert-Ana, A. Sankarankutty, P. Marques, and J. Camacho-Guerrero, “Prototyping a capture and access application to document medical grand rounds,” in *Anais do XIII Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, (Gramado - Brazil), p. 8p., Available on ACM Digital Library, Oct. 2007.
- [5] R. d. F. B. Neto, J. A. C. Guerrero, Á. Barreiro, J. Parapar, and A. A. Macedo, “An automatic linking service of document images reducing the effects of ocr errors with latent semantics,” in *Proc. of the 25th Annual ACM Symposium on Applied Computing (ACM SAC’10)*, pp. 13–17, 2010.
- [6] R. d. F. B. Neto, J. A. C. Guerrero, Á. Barreiro, J. Parapar, and A. A. Macedo, “The use of latent semantic indexing to mitigate ocr effects of related document images,” *Journal of Universal Computer Science (Print)*, vol. 17, pp. 64–80, 2011.
- [7] H. C. Pessotti, P. M. A. Marques, and A. A. Macedo, “Framework para classificação automática de tomografias computadorizadas de alta resolução para auxílio ao diagnóstico de lesões de pulmão,” in *Anais do XXI Congresso de Engenharia Biomédica (CBEB 2008)*, (Salvador - Brazil), pp. 621–624, November 2008.
- [8] H. C. Pessotti, L. O. M. Junior, E. G. Soares, and A. A. Macedo, “Freds: Framework para redução da descontinuidade semântica em imagens médicas,” in *Workshop de Informática Médica (WIM) - XXXI Congresso da Sociedade Brasileira de Computação*, pp. 1782–1791, 2011.
- [9] A. A. Macedo, H. Pessotti, L. F. Almansa, J. C. Felipe, and E. Kimura, “Morphometric information to reduce the semantic gap in the characterization of microscopic images of thyroid nodules,” *Computer Methods and Programs in Biomedicine*, p. 22p., 2016.
- [10] M. G. Manzato, A. A. Macedo, and R. Goularte, “Evaluation of video news classification techniques for automatic content personalization,” *Int. J. Advanced Media and Communication*, vol. 3, no. 4, pp. 383–403, 2009.

- [11] J. T. Pollettini, S. R. G. Panico, J. C. Daneluzzi, R. Tinós, J. A. Baranauskas, and A. A. Macedo, “Using machine learning classifiers to assist healthcare-related decisions: Classification of electronic patient records,” *Journal of Medical Systems*, vol. 36, pp. 3861–3874, 2012.
- [12] J. T. Pollettini, J. A. Baranauskas, E. S. Ruiz, M. d. G. C. Pimentel, and A. A. Macedo, “Surveillance for the prevention of chronic diseases through information association,” *BMC Medical Genomics*, vol. 7, p. 11p., jan 2014.
- [13] A. A. Macedo, J. Polettini, J. A. Baranauskas, and J. Chaves, “A health surveillance software framework to design the delivery of information on preventive healthcare strategies,” *Submitted with minor revisions to the Journal of Biomedical Informatics*, p. 22p., 2016.
- [14] A. A. Macedo, M. d. G. Pimentel, and J. Polettini, “A software framework for health surveillance recommender systems,” in *Submitted to the Proceedings of the DocEng2016*, p. 4p., 2016.
- [15] L. F. Alamnsa and A. A. Macedo, “Sistema de informação para perguntas e respostas em doenças crônicas,” in *To be published to the XVI Medical Informatics Workshop - CSBC 2016*, (Porto Alegre/RS - Brazil), p. 10p., July 2016.
- [16] V. Bush, “As we may think,” *Atlantic Monthly*, vol. 176, pp. 101 – 108, July 1945.
- [17] D. C. Engelbart, “Augmenting society’s collective IQ.” Keynote Abstract - Hypertext 2004, 2004. <http://www.doungelbart.org/pubs/augment-133319.html>.
- [18] World Wide Web Consortium, “World Wide Web Consortium.” Visited in May 2016. Online in World Wide Web, 2016. <http://www.w3.org>.
- [19] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY: Addison Wesley, 1999.
- [20] J. Cowie and W. Lehnert, “Information extraction,” *Communications of the ACM*, vol. 39, no. 1, pp. 80–91, 1996.
- [21] R. Grishman, “Information extraction: Techniques and challenges,” in *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE ’97, (London, UK), pp. 10–27, Springer-Verlag, 1997.
- [22] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. 2007.
- [23] K. B. Cohen and L. Hunter, “Getting started in text mining,” vol. 4, pp. 1–20, 2008.
- [24] A. A. Macedo, M. G. C. Pimentel, and J. A. C. Guerrero, “Latent semantic linking over homogeneous repositories,” in *Proceedings of the ACM Symposium on Document Engineering (DocEng)*, (Atlanta, GA, USA), pp. 144–151, ACM Press, November 2001.
- [25] A. A. Macedo, M. G. C. Pimentel, and J. A. Cammacho-Guerrero, “An infrastructure for open latent semantic linking,” in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, (College Park, Maryland, USA), pp. 107–116, ACM Press, 2002.

- [26] A. A. Macedo, K. N. Truong, J. A. Camacho-Guerrero, and M. G. C. Pimentel, “Automatically sharing Web experiences through a hyperdocument recommender system,” in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, (Nottingham, UK), pp. 48–56, ACM Press, August 2003.
- [27] A. A. Macedo, J. A. Camacho-Guerrero, R. G. Cattelan, V. R. Inácio Jr, and M. G. C. Pimentel, “Interaction alternatives for linking everyday presentations,” in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, (Santa Cruz, CA, USA), pp. 112–113, ACM Press, August 2004.
- [28] A. A. Macedo, J. A. Camacho-Guerrero, and M. d. G. C. Pimentel, “Bilingual linking service for the web,” in *Proc. of 12th Symposium on String Processing and Information Retrieval (SPIRE)* (S.-V. B. Heidelberg, ed.), vol. 3772, (Buenos Aires - Argentina), pp. 45–48, Lecture Notes in Computer Science (LNCS), Oct. 2005.
- [29] A. A. Macedo, J. A. C. Guerrero, R. G. Cattelan, V. d. R. I. Jr, and M. d. G. C. Pimentel, “Multimodal interactions for linking everyday presentations in a ubiquitous computing infrastructure,” in *WebMedia 2010*, pp. 155–162, SBC, 2010.
- [30] A. A. Macedo, J. T. Pollettini, and E. V. Munson, “A chronic illness system using biomedical knowledge sources and relevance feedback,” in *IEEE 28th International Symposium on Computer-Based Medical Systems (IEEE-CBMS)*, pp. 244–249, IEEE, 2015.
- [31] H. Small, “Co-citation in the scientific literature: A new measure of the relationship between two documents,” *Journal of the American Society for Information Science*, vol. 24, pp. 265 – 269, February 1973.
- [32] M. Agosti and F. Crestani, “A methodology for the automatic construction of a hypertext for information retrieval,” in *Proceedings of ACM/SIGAPP Symposium on Applied Computing: states of the art and practice*, (Indianapolis, IN, USA), pp. 745 – 753, ACM Press, 1993.
- [33] G. Salton and J. Allan, “Selective text utilization and text traversal,” in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, (Seattle, Washington, USA), pp. 131 – 144, ACM Press, 1993.
- [34] J. Allan, “Automatic hypertext link typing,” in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, pp. 42 – 52, ACM Press, 1996.
- [35] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” *Computer Network and ISDN System*, vol. 30, no. 1 – 7, pp. 107 – 117, 1998.
- [36] J. M. Kleinberg, “Authoritive sources in a hyperlinked environment,” *Journal of the ACM*, vol. 46, no. 5, pp. 604 – 632, 1999.
- [37] S. J. Green, “Building hypertext links by computing semantic similarity,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, pp. 713 – 730, Sempember 1999.
- [38] N. Wacholder, D. K. Evans, and J. L. Klavans, “Automatic identification and organization of index terms for interactive browsing,” in *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, (Roanoke, Virginia, United States), pp. 126–134, ACM Press, 2001.

- [39] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura, “Refinement of TF-IDF schemes for Web pages using their hyperlinked neighboring pages,” in *Proceedings of the ACM Conference on Hypermedia and Hypertext*, (Nottingham, UK), pp. 198 – 207, ACM Press, 2003.
- [40] K. Chen and L. Liu, “ivibrate: Interactive visualization-based framework for clustering large datasets,” *ACM Trans. Inf. Syst.*, vol. 24, no. 2, pp. 245–294, 2006.
- [41] S. Soderland and B. Mandhani, “Moving from textual relations to ontologized relations.,” in *AAAI Spring Symposium: Machine Reading*, pp. 85–90, AAAI, 2007.
- [42] K. Qian, S. Hirokawa, K. Ejima, and X. Du, “A fast associative mining system based on search engine and concept graph for large-scale financial report texts,” in *Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on*, pp. 675–679, Sept 2010.
- [43] A. Fader, S. Soderland, and O. Etzioni, “Identifying relations for open information extraction,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11*, (Stroudsburg, PA, USA), pp. 1535–1545, Association for Computational Linguistics, 2011.
- [44] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, “Semantic path based personalized recommendation on weighted heterogeneous information networks,” in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM ’15*, (New York, NY, USA), pp. 453–462, ACM, 2015.
- [45] G. D. Abowd, “Classroom 2000: an experience with the instrumentation of a living educational environment,” *IBM Systems Journal*, vol. 38, pp. 508 – 530, 1999.
- [46] M. Guzdial, “Supporting learners as users,” *The Journal of Computer Documentation*, vol. 23, no. 2, pp. 3 – 13, 1999.
- [47] M. G. C. Pimentel, A. A. Macedo, and G. D. Abowd, “Linking homogeneous web-based repositories,” in *Proceedings of International Workshop on Information Integration on the Web*, (Rio de Janeiro-RJ,Brazil), pp. 35 – 42, April 2001. <http://www.cos.ufrj.br/wiiw/schedule.html>.
- [48] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, “Using latent semantic analysis to improve access to textual information,” in *Conference Proceedings on Human Factors in Computing Systems*, pp. 281 – 285, 1998.
- [49] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Stretter, and K. E. Lochbaum, “Information retrieval using a singular value decomposition model of latent semantic structure,” in *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, (Grenoble, France), pp. 465 – 480, ACM Press, 1988.
- [50] J. A. Camacho-Guerrero, A. A. Macedo, and R. P. M. Fortes, “Uma infra-estrutura configurável para serviços de criação automática de ligações,” in *Anais do VII Brazilian Symposium on Multimedia and Hypermedia System (SBMidia)*, (Fortaleza-CE,Brazil), pp. 298 – 305, 2002.

- [51] A. A. Macedo, J. A. Camacho-Guerrero, and M. G. C. Pimentel, “Incluindo abordagens de recuperação de informação em serviços de criação de hiperligações,” in *XXVIII Conferencia Latinoamericana de Informática*, (Montevideo, Uruguai), p. 8p. (Eletronically Published), November 2002.
- [52] R. d. F. Bulcão Neto, A. A. Macedo, J. A. C. Guerrero, and M. d. G. C. Pimentel, “Configurable semantic services leveraging context-aware applications,” in *Anais do XI Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, (Poços de Caldas - Brazil), pp. 1–9, SBC, Nov. 2005.
- [53] D. Roberts and R. Johnson, “Evolving Frameworks: A Pattern Language for Developing Object-Oriented Frameworks,” in *Proceedings of the Third Conference on Pattern Languages and Programming*, vol. 3, Addison-Wesley, 1996.
- [54] A. A. Macedo, “Recuperação de informações e aplicações web,” in *Tópicos em Tecnologias Web & Multimídia (Capítulo 4)*, pp. 147–196, Ribeirão Preto, SP, Brazil: SBC, 2004.
- [55] E. Oz, *Management Information Systems*, ch. Capter 9 - Chanllenges of Global Information Systems, p. 315. sixth edition – first published 1997 ed., 2009.
- [56] J. A. Camacho-Guerrero and A. A. Macedo, “Tecnologias de recuperação de informações na web,” in *Web e Multimídia: Desafios e Soluções (Capítulo 6)*, pp. 167–196, 2005.
- [57] J. J. Rocchio, “Relevance feedback in information retrieval,” in *The Smart Retrieval System - Experiments in Automatic Document Processing* (G. Salton, ed.), 1971.
- [58] A. Macedo, J. Camacho-Guerrero, E. Ruiz, and L. Pellozo, “Criação automática de hiperligações semânticas a partir da categorização vetorial por definição de centróides,” in *Anais do XIII Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, (Gramado - Brazil), p. 8p., Available on ACM Digital Library, Oct. 2007.
- [59] G. Miller, “Wordnet: An on-line lexical database,” *International journal of lexicography*, vol. 3, no. 4, pp. 235 – 312, 1990.
- [60] J. A. Camacho-Guerrero, A. A. Carvalho, and A. A. Macedo, “Categorização de informações no suporte a identificação automática de hiperligações semânticas,” in *Anais do XII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, (Natal - Brazil), p. 4p., 2006.
- [61] J. A. Camacho-Guerrero, A. A. Carvalho, M. d. G. C. Pimentel, E. V. Munson, and A. A. Macedo, “Clustering as an approach to support the automatic definition of semantic hyperlinks,” in *Proc. of The Eighteenth ACM Conference on Hypertext and Hypermedia (HT’07)*, (Manchester - UK), pp. 81–84, ACM Press, Sept. 2007.
- [62] T. M. da Costa, P. S. Perez, M. P. da Silva, and A. A. Macedo, “Relato da experiência de desenvolvimento de um sistema de informação para acompanhamento da saúde e desenvolvimento humano de recém-nascidos em uma unidade de atendimento primário,” in *Poster publicado no 13o Simpósio Internacional de Iniciação Científica da USP*, (Ribeirão Preto - SP), p. 1p., 2005.

- [63] L. S. Gugelmim and A. A. Macedo, "Identificação automática de relacionamentos entre informações similares advindas da documentação de projeto de software," in *Anais do XII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, (Natal - Brazil), p. 4p., 2006.
- [64] M. Santos, C. Jomoli, R. Nazaré, W. Zhuofan, and A. Macedo, "Estudo comparativo de algoritmos e estruturas de dados para indexação de grandes volumes de informações textuais," in *Anais do XI Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, (Poços de Caldas - Brazil), p. 3p., Nov. 2005.
- [65] M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, *Current Challenges in Patent Information Retrieval*, vol. 29. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2011. <http://www.springerlink.com/content/978-3-642-19231-9#section=878399&page=1>.
- [66] Agência USP de Inovação, "Patentes." Internet (Visited: 21/03/2016), 2016. <http://inovacao.usp.br/propriedade-intelectual/patentes/>.
- [67] T. G. O. Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, May 2000. <http://dx.doi.org/10.1038/75556>.
- [68] M. B. Dutra, J. A. C. Guerrero, J. A. Baranauskas, and A. A. Macedo, "Busca guiada de patentes de bioinformática," in *XIII Workshop de Informática Médica (WIM) - XXXIII Congresso da Sociedade Brasileira de Computação (CSBC)*, (Maceió), pp. 150–159, Sociedade Brasileira de Computação, 2013.
- [69] Y. Rogers and M. Scaife, *How can interactive multimedia facilitate learning?* AAI, 1998.
- [70] M. Turk and G. Robertson, "Perceptual user interfaces (introduction)," *Commun. ACM*, vol. 43, pp. 32–34, Mar. 2000.
- [71] C. Baber and B. Mellor, "Using critical path analysis to model multimodal human-computer interaction," *Int. J. Hum.-Comput. Stud.*, vol. 54, pp. 613–636, April 2001.
- [72] M. Weiser, "The computer for the 21st century," *Scientific American*, vol. 265, pp. 94 – 104, September 1991. <http://www.ubiq.com/hypertextweiserSciAmDraft3.html>.
- [73] R. d. F. Bulcão Neto, J. A. Camacho-Guerrero, P. M. A. Marques, Lucas, and A. A. Macedo, "Um arcabouço de software para apoio ao desenvolvimento de aplicações médicas com tecnologias de computação ubíqua," in *Anais do VII Workshop de Informática Médica*, (Porto de Galinhas - Brazil), p. 4p., SBC, Jul. 2007.
- [74] R. d. F. Bulcão Neto, J. A. Camacho-Guerrero, and A. A. Macedo, "Um arcabouço de software para apoio ao desenvolvimento de aplicações de apoio ao diagnóstico médico," in *Poster publicado no Congresso Tecnologia e Humanização na Comunicação em Saúde (ConTIC-Saúde)*, (Ribeirão Preto - SP), p. 1p., 2007.
- [75] R. de Freitas Bulcão-Neto, J. A. Camacho-Guerrero, and A. A. Macedo, "A prototype documenter system for medical grand rounds," in *Proceeding of the eighth ACM Symposium on Document Engineering (DocEng)*, (New York, NY, USA), pp. 104–105, ACM, 2008.

- [76] W. Sharrock and R. Anderson, *Epistemology: Professional Scepticism*, ch. Ethnomethodology and the Human Sciences, pp. 51–76. Cambridge University Press, 1991.
- [77] R. d. F. Bulcão-Neto, A. A. Macedo, L. Wichert-Ana, A. Sankarankutty, P. M. A. Marques, and J. A. Camacho-Guerrero, “Supporting ethnographic studies of ubiquitous computing in the medical grand round experience,” in *23rd Annual ACM Symposium on Applied Computing (ACM SAC)*, (Fortaleza - CE - Brazil), pp. 1641–1645, 1641-1645, Nov. 2008.
- [78] R. F. Bulcão-Neto, J. A. Camacho-Guerrero, and A. A. Macedo, “Automatic documentation of users interactions with dicom images: A case study in medical grand rounds,” in *Anais do VIII Workshop de Informática Médica (Evento paralelo ao XXVIII Congresso da Sociedade Brasileira de Computação)*, (Belém - Brazil), p. 4p., July 2008.
- [79] R. F. Bulcão-Neto, J. A. Camacho-Guerrero, and A. A. Macedo, “Extension of captured information in pervasive healthcare systems: A case study,” in *Proc. of IEEE 11th International Conference on Computational Science and Engineering (International Symposium on Scientific and Engineering Computing (SEC-08))*, (São Paulo - Brazil), pp. 19–24, Available on IEEE Digital Library, July 2008.
- [80] J. A. C. Guerrero and A. A. Macedo, “Explorando HTML5, CSS3 e JQueryMobile no controle e monitoramento de casas inteligentes,” in *Tópicos em Multimídia, Hipermídia e Web*, pp. 40–74, Bahia: Sociedade Brasileira de Computação, 1 ed., 2013.
- [81] The Clendening Library Group, “Digital Clendening.” Internet (Visited: 10/03/2016), 2016. <http://clendening.kumc.edu>.
- [82] Google Corp., “Google Patent Search Homepage.” Internet (Visited: 10/03/2016), 2016. <http://patents.google.com>.
- [83] Proin, “Arquivo Publico do Estado e Universidade de Sao Paulo.” Internet (Visited: 10/03/2016), 2016. <http://www.usp.br/proin>.
- [84] Obras Raras, “Obras Raras.” Internet (Visited: 10/03/2016), 2016. <http://www.obrasraras.usp.br>.
- [85] K. Taghva, J. Borsack, and A. Condit, “Effects of OCR errors on ranking and feedback using the vector space model,” *Information Processing Management*, vol. 32, no. 3, pp. 317–327, 1996.
- [86] A. Depeursinge, H. Müller, A. Hidki, P.-A. Poletti, A. Platon, and A. Geissbuhler, “Image-based diagnostic aid for interstitial lung disease with secondary data integration,” vol. 6514, pp. 65143P–65143P, 2007.
- [87] P. G. I. I. e. a. Azevedo-Marques, P. M.; Tahoces, “Automated characterization of interstitial lung disease on postero-anterior chest radiographs,” *Int. Journal of Computer Assisted Radiology and Surgery*, vol. 1, pp. 350–352, 2006.
- [88] Y. Uchiyama, S. Katsuragawa, H. Abe, J. Shiraishi, F. Li, Q. Li, C. T. Zhang, K. Suzuki, and K. Doi, “Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography,” *Medical Physics*, vol. 30, pp. 2440–2454, 9 2003.

- [89] P. Marques, A. Macedo, H. Pessotti, S. Salomão, A. Traina, and M. da Silva, "CAD-PACS integration: a framework for pattern recognition of diffuse lung disease in HRCT," in *Proc. of Computer Assisted Radiology and Surgery - Joint Congress of CAR / ISCAS / CMI / CAD / EuroPACS / CURAC, 2009, Berlin. Joint Congress of CAR / ISCAS / CMI / CAD / EuroPACS / CURAC*, (Berlin - Germany), pp. S.180–S.181, June 2009.
- [90] A. L. S. Maia, L. S. Ward, G. A. d. Carvalho, H. Graf, R. M. d. B. Maciel, L. M. Z. Maciel, P. W. Rosário, and M. Vaisman, "Nódulos de tireóide e câncer diferenciado de tireóide: consenso brasileiro," *Arq Bras Endocrinol Metab*, vol. 51, no. 5, pp. 867–893, 2007.
- [91] A. Mueen, R. Zainuddin, and M. S. Baba, "Automatic multilevel medical image annotation and retrieval.," *J. Digital Imaging*, vol. 21, no. 3, pp. 290–295, 2008.
- [92] D. F. d. Carvalho, J. A. Camacho-Guerrero, P. M. d. A. Marques, and A. A. Macedo, "Lyria PACS: A case study saves ten million dollars in a brazilian hospital," in *IEEE 28th International Symposium on Computer-Based Medical Systems (IEEE-CBMS)*, pp. 326–329, 2015.
- [93] R. F. Salles, *Análise de um Programa de intervenção com bebês e famílias atendidas em unidades básicas de saúde - SUS*. 2001.
- [94] R. F. Salles, "Análise de um programa de intervenção com bebês e famílias atendidas em unidades de saúde – SUS," dissertação de mestrado, programa de pós-graduação em educação especial, Universidade Federal de São Carlos, São Carlos, São Paulo, Brasil, 2000.
- [95] S. R. G. Panico, M. L. Canziani, and N. Guerchon, *Indicadores Nipe: Subsídios para Políticas Municipais de Saúde*, vol. 1, ch. Políticas Públicas Municipais. São Carlos-SP: NIPE, 1 ed., 1997.
- [96] D. S. de Paula, S. Panico, J. C. Daneluzzi, J. C. F. E. E. S. Ruiz, and A. A. Macedo, "Sistema de informação de apoio ao programa de educação para pais e famílias," in *Anais do XI Congresso Brasileiro de Informática em Saúde (CBIS08). Electronically Published.*, (Campos do Jordão - Brazil), p. 6p., December 2008.
- [97] J. T. Pollettini, R. Tinós, S. Panico, J. C. Daneluzzi, and A. A. Macedo, "Classificação automática de pacientes para atendimento médico pediátrico multidisciplinar a partir do seu grau de vigilância," in *Anais do VIII Workshop de Informática Médica (Evento paralelo ao XXVIII Congresso da Sociedade Brasileira de Computação)*, (Belém - Brazil), pp. 61–70, SBC, July 2008.
- [98] J. T. Pollettini, G. H. B. Miranda, R. Goularte, S. Panico, J. C. Daneluzzi, and A. A. Macedo, "Sistema de informação geográfica: uma abordagem integrada a sistemas de informação em saúde," in *Anais do XII Congresso Brasileiro de Informática em Saúde (CBIS)*, p. 6p., 2010.
- [99] J. T. Pollettini, F. P. Nicolas, S. Panico, J. C. Daneluzzi, R. Tinós, J. A. Baranauskas, and A. A. Macedo, "A software architecture-based framework supporting suggestion of medical surveillance level from classification of electronic patient records," in *Proc of the 12th IEEE International Conference on Computational Science and Engineering*, (Vancouver - Canadá), pp. 166–173, IEEE Computer Society, August 2009.

- [100] J. T. Pollettini, R. Tinós, S. Panico, J. C. Daneluzzi, and A. A. Macedo, “Vigilância em atenção básica a saúde a partir do uso de relevance feedback para classificação de pacientes em diferentes níveis de cuidado em saúde,” in *Anais do IX Workshop de Informática Médica (Evento paralelo ao XXIX Congresso da Sociedade Brasileira de Computação)*, (Bento Gonçalves - RS), pp. 1945–1954, SBC, July 2009.
- [101] National Library of Medicine (US)., Bethesda, MD, USA, *UMLS Reference Manual [Internet]*, 1999.
- [102] F. P. Nicolas and A. A. Macedo, “Mecanismos de classificação para definição automática de medidas que identificam pessoas requerendo diferentes graus de atendimento médico: um estudo de caso usando umls,” in *Poster publicado no 17o Simpósio Internacional de Iniciação Científica da USP*, (Ribeirão Preto - SP), p. 1p., 2009.
- [103] F. P. Nicolas, J. T. Pollettini, S. Panico, J. C. Daneluzzi, E. E. R. Seron, and A. A. Macedo, “Determinação e análise de tipos semânticos usando umls e árvores de decisão para aprimorar a determinação automática do grau de vigilância de pacientes,” in *Anais do XII Congresso Brasileiro de Informática em Saúde (CBIS)*, p. 6p., 2010.
- [104] F. P. Nicolas, J. T. Pollettini, S. Panico, J. C. Daneluzzi, and A. A. Macedo, “O uso de umls para aprimorar a recomendação de graus de vigilância para pacientes do setor primário,” in *Workshop de Informática Médica (WIM) - XXX Congresso da Sociedade Brasileira de Computação*, p. 10p., 2010.
- [105] “Cardiovascular diseases (cvds), fact sheet nº317, updated march 2013.” <http://www.who.int/mediacentre/factsheets/fs317/en/index.html>.
- [106] A. Verma, N. Kasabov, E. Rush, and Q. Song, “Ontology Based Personalized Modeling for Chronic Disease Risk Analysis: An Integrated Approach,” in *Advances in Neuro-Information Processing* (M. Koeppen, N. Kasabov, and G. Coghill, eds.), vol. 5506 of *Lecture Notes in Computer Science*, pp. 1204–1210, Berlin - Heidelberg: Springer, 2009.
- [107] “10 facts about chronic diseases.” World Health Organization. Internet (Visited: 31/03/2016). <http://www.who.int/features/factfiles/chp/en/index.html>.
- [108] “10 facts about chronic diseases: Fact 10.” World Health Organization. Internet (Visited: 31/03/2016). http://www.who.int/features/factfiles/chp/10_en.html.
- [109] D. J. P. Barker, “Fetal and infant origins of adult disease,” *Monatsschrift Kinderheilkunde*, vol. 149, pp. S2–S6, Jun 2001.
- [110] J. J. Pasternak, *Uma Introdução a Genética Molecular Humana: Mecanismos das Doenças Hereditárias*. Guanabara Koogan S.A., 2 ed., jun 2007.
- [111] D. Barker, “The fetal and infant origins of adult disease,” *BMJ*, vol. 301, p. 1111, July 1990.
- [112] D. Barker, “Fetal origins of cardiovascular disease,” *Ann Med*, vol. Suppl 1, pp. 3–6, Apr 1999.
- [113] P. D. Wadhwa, C. Buss, S. Entringer, and J. M. Swanson, “Developmental origins of health and disease: brief history of the approach and current focus on epigenetic mechanisms,” *Seminars in reproductive medicine*, vol. 27, pp. 358–68, Sept. 2009.

- [114] M. W. Gillman, D. Barker, D. Bier, F. Cagampang, J. Challis, C. Fall, K. Godfrey, P. Gluckman, M. Hanson, D. Kuh, P. Nathanielsz, P. Nestel, and K. L. Thornburg, "Meeting report on the 3rd International Congress on Developmental Origins of Health and Disease (DOHaD).," *Pediatric research*, vol. 61, pp. 625–9, May 2007.
- [115] R. W. Fogel, "Second Thoughts on the European Escape from Hunger: Famines, Chronic Malnutrition, and Mortality Rates," in *Nutrition and Poverty* (S. R. Osmani, ed.), pp. 243–86, New York: Oxford University Press Clarendon Press, 1992.
- [116] "Developmental origins of health and disease (DOHaD)." DOHaD Society, Internet (Visited: 31/03/2016). <https://dohadsoc.org>.
- [117] "A portuguese stop word list."
- [118] J. T. Pollettini and A. A. Macedo, "Poster: Perspectives on the use of bioinformatics information to help healthcare professionals on chronic diseases prevention," in *Proc. of the 5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (XMeeting)*, p. 1p., 2009.
- [119] J. T. Pollettini, B. Gisele Miranda, R. Goularte, S. Panico, J. C. Daneluzzi, and A. A. Macedo, "Sistema de Informação Geográfica: uma Abordagem Integrada a Sistemas de Informação em Saúde," in *Congresso Brasileiro de Informática em Saúde (CBIS)*, Jan. 2010.
- [120] J. T. Pollettini and A. A. Macedo, "Poster: Chronic disease prevention: A translational bioinformatics approach," in *2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCBABS)*, p. 258, IEEE, 2011.
- [121] S. Bird, E. Klein, and E. Loper, *NLTK Book*. Sebastopol, CA: O'Reilly Media, 2009.
- [122] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in *Proceedings of the AMIA Annual Symposium: 3-7 November 2001; Washington, DC* (S. Bakken, ed.), pp. 17–21, Stoneham: Butterworth-Heinemann American Medical Informatics Association, 2011.
- [123] J. Chaves, J. Pollettini, and A. Macedo, "Relating biomedical information using information mapping supported by semantic web," in *Proceedings of the 15th World Congress on Health and Biomedical Informatics*, MEDINFO 2015, p. 1p., 2015.
- [124] "RDFLib: rdf library for python, including a sparql implementation." W3C na Internet (Visited: 31/03/2016). <https://www.w3.org/2001/sw/wiki/RDFLib> ou <https://github.com/RDFLib/rdfliib>.
- [125] "OGC Open Geospatial Consortium: Making location count." Open GeoSpatial Org in Internet (Visited: 31/03/2016). <http://www.opengis.net/doc/IS/geosparql/1.0>.
- [126] "Virtuoso Universal Server." OpenLink Software in Internet (Visited: 31/3/2016). <http://virtuoso.openlinksw.com/>.
- [127] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, pp. 211–220, 2012.

- [128] N. P. Er and I. Cicekli, “A Factoid Question Answering System Using Answer Pattern Matching,” in *International Joint Conference on Natural Language Processing*, (Nagoya, Japan), pp. 854–858, 2013.
- [129] C. Monz, *From document retrieval to question answering*. The address of the publisher: Amsterdam: Institute for Logic, Language and Computation, 2003.
- [130] D. Zhang and W. S. Lee, “Question classification using support vector machines,” in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, (New York, NY, USA), pp. 26–32, ACM, 2003.
- [131] P. Gupta and V. Gupta, “A survey of text question answering techniques,” *International Journal of Computer Applications*, vol. 53, no. 4, pp. 1–8, 2012.
- [132] T. Strachan and A. P. Read, *Human Molecular Genetics*. Garland Science, 2 ed., 1999.
- [133] P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma, “MedMeSH Summarizer: Text Mining for Gene Clusters,” in *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*, p. 18, Society for Industrial and Applied Mathematics, SIAM, 2002.
- [134] A. KOSCHMIEDER, K. ZIMMERMANN, S. TRISSL, T. STOLTMANN, and U. LEISER, “Tools for managing and analyzing microarray data,” *Briefings in bioinformatics*, vol. 13, no. 1, pp. 46–60, 2012.
- [135] M. B. Mohan, *Computational Genomics: Theory and Application*, ch. An introduction to microarray data analysis, pp. 225–249. Horizon Scientific Press, Norwich, UK, 2004.
- [136] M. C. Monard and J. A. Baranauskas, *Sistemas Inteligentes: Fundamentos e Aplicações*, ch. Conceitos sobre aprendizado de máquina. Editora Manole Ltda, 2003.
- [137] A. Splendore, “Para que existem as regras de nomenclatura genética?,” *Revista Brasileira de Hematologia e Hemoterapia*, vol. 27, pp. 148–152, 06 2005.
- [138] D. S. d. Paula and A. A. Macedo, “Poster: Textual summarization in bioinformatics domain,” in *Proc. of the 5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology*, 2009.
- [139] D. S. d. Paula and A. A. Macedo, “Método para auxiliar a interpretação de clusters de expressão gênica considerando sumarização automática,” in *XII Workshop de Informática Médica (WIM) - XXXII Congresso da Sociedade Brasileira de Computação*, 2012.
- [140] I. Gurol-Urganci, T. de Jongh, V. Vodopivec-Jamsek, J. Car, and R. Atun, “Mobile phone messaging for communicating results of medical investigations.,” *The Cochrane database of systematic reviews*, vol. 6, pp. CD007456–CD007456, 2011.
- [141] “3 em cada 10 brasileiros são donos de smartphones.” *Revista Exame na Internet* (Visited: 31/03/2016), 2013. <http://exame.abril.com.br/tecnologia/noticias/3-em-cada-10-brasileiros-sao-donos-de-smartphones>.

- [142] “Demografia médica 2011.” Conselho Regional de Medicina do Estado de São Paulo na Internet (Visited: 31/03/2016). www.cremesp.org.br/?siteAcao=CentroDados&acao=detalhes_capitulos&cod_capitulo=4.
- [143] L. X. X. WANG, X.; RYAN, “Association between air pollution and low birth weight: a community-based study,” *Environmental Health Perspective*, vol. 105, pp. 514–520, 1997.
- [144] C. LEE, K.; GAY, “Sleep in late pregnancy predicts length of labor and type of delivery,” *American Journal of Obstetrics and Gynecology*, vol. 191, pp. 2041–2046, 2004.
- [145] “What are the factors that put pregnancy at risk?.” NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT na Internet (Visited: 31/03/2016). <https://www.nichd.nih.gov/health/topics/high-risk/conditioninfo/Pages/factors.aspx>.
- [146] “OMS: Brasil reduz mortalidade materna em 43% de 1990 a 2013.” Portal Brasil na Internet (Visited: 31/03/2016). <http://www.brasil.gov.br/saude/2014/05/oms-brasil-reduz-mortalidade-materna-em-43-de-1990-a-2013>.
- [147] S. B. M. C. M. E. F. MORSE, M.; FONSECA, “Mortalidade neonatal: Características assistências e biológicas dos recém-nascidos e de suas mães,” *Caderno de Saúde Pública*, vol. 27, pp. 623–638, abril 2011.
- [148] J. Kim and T. Wilhelm, “What is a complex graph?,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 11, pp. 2637–2652, 2008.
- [149] D. J. Watts and S. H. Strogatz, “Collective dynamics of /‘small-world/’ networks,” *Nature*, vol. 393, pp. 440–442, 06 1998.
- [150] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509–512, 1999.
- [151] M. E. Newman, “The structure and function of complex networks,” *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.
- [152] J. Tang, H. Gao, X. Hu, and H. Liu, “Exploiting homophily effect for trust prediction,” in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 53–62, ACM, 2013.
- [153] Y. Sun and J. Han, “Mining heterogeneous information networks: a structural analysis approach,” *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 20–28, 2013.
- [154] B. E. S. Youssef, *Models for the Generation of Heterogeneous Complex Networks*. PhD thesis, 2015.
- [155] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, “A survey of heterogeneous information network analysis,” *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 2015.
- [156] Z. M. Z. B. Long and P. S. Yu, “Co-clustering by block value decomposition,” in *Proceedings of the KDD*, p. 635–640, 2005.

- [157] M. Jamali and L. Lakshmanan, “Heteromf: recommendation in heterogeneous information networks using context dependent factor models,” in *Proceedings of the WWW*, p. 643–654, 2013.
- [158] P. Z. Z. Y. H. C. Y. Sun, J. Han and T. Wu, “Rankelus: integrating clustering with ranking for heterogeneous information network analysis,” in *Proceeding of the EDBT*, p. 565–576, 2009.
- [159] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, “Mathematical formulation of multilayer networks,” *Physical Review X*, vol. 3, no. 4, p. 041022, 2013.
- [160] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, “Multilayer networks,” *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.
- [161] Y. Sun and J. Han, “Mining heterogeneous information networks: principles and methodologies,” *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 1–159, 2012.
- [162] X. F. Wang and G. Chen, “Complex networks: Small-world, scale-free and beyond,” *IEEE Circuits and Systems Magazine*, vol. 3, pp. 6–20, September 2003.
- [163] D. R. Amancio, *Classificação de textos com redes complexas*. PhD thesis, Universidade de São Paulo, 2013.
- [164] H. Ma, I. King, and M. R. Lyu, “Learning to recommend with social trust ensemble,” in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’09, (New York, NY, USA), pp. 203–210, ACM, 2009.
- [165] X. Yang, H. Steck, and Y. Liu, “Circle-based recommendation in online social networks,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’12, (New York, NY, USA), pp. 1267–1275, ACM, 2012.
- [166] C. Yang, J. Sun, J. Ma, S. Zhang, G. Wang, and Z. Hua, “Scientific collaborator recommendation in heterogeneous bibliographic networks,” in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pp. 552–561, Jan 2015.
- [167] P. Zhao, J. Han, and Y. Sun, “P-rank: a comprehensive structural similarity measure over information networks,” in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 553–562, ACM, 2009.
- [168] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “Pathsim: Meta path-based top-k similarity search in heterogeneous information networks,” *VLDB’11*, vol. 4, no. 11, p. 12, 2011.
- [169] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, “Graph regularized transductive classification on heterogeneous information networks,” in *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, ECML PKDD’10, (Berlin, Heidelberg), pp. 570–586, Springer-Verlag, 2010.

- [170] Y. Zhou and L. Liu, “Activity-edge centric multi-label classification for mining heterogeneous information networks,” in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’14, (New York, NY, USA), pp. 1276–1285, ACM, 2014.
- [171] S. D. Chen, Y. Y. Chen, J. Han, and P. Moulin, “A feature-enhanced ranking-based classifier for multimodal data and heterogeneous information networks,” in *2013 IEEE 13th International Conference on Data Mining*, pp. 997–1002, Dec 2013.
- [172] J. A. Baranauskas, O. P. N. Sérgio, R. Nozawa, and A. A. Macedo, “A tree-based algorithm for attribute selection,” *Submitted with major revisions to the Applied Intelligence*.
- [173] P. S. Perez, S. R. NOZAWA, A. A. Macedo, and J. A. BARANAUSKAS, “Windowing improvements towards more comprehensible models,” *Knowledge-Based Systems*, vol. 92, pp. 9–22, 2016.
- [174] E. A. Tanaka, S. R. Nozawa, A. A. Macedo, and J. A. Baranauskas, “A multi-label approach using binary relevance and decision trees applied to functional genomics,” *Journal of Biomedical Informatics*, vol. 54, pp. 85–95, 2015.
- [175] P. S. Perez, A. H. Bevilacqua, A. Ghelfi, A. A. Macedo, S. R. Nozawa, and J. A. Baranauskas, “A software tool for information management and data mining of biological data for studying adaptation of living organisms in amazonia,” in *Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCBBI)*, p. 7p., 2011.
- [176] O. Picchi, A. A. Macedo, P. M. d. A. Marques, and J. A. Baranauskas, “Uma metodologia para estruturação de laudos médicos usando ontologias,” in *Workshop de Informática Médica (WIM) - XXXI Congresso da Sociedade Brasileira de Computação*, pp. 1816–1825, SBC, 2011.
- [177] O. Picchi, S. R. Nozawa, R. A. R. Mitrowsky, A. A. Macedo, and J. A. Baranauskas, “Applying decision trees to gene expression data from dna microarrays: A leukemia case study,” in *Workshop de Informática Médica (WIM) - XXX Congresso da Sociedade Brasileira de Computação*, p. 10p., 2010.