ROBERTO SILVA BAPTISTA

MAPA DE INTERESSES DE AUTORES DE PUBLICAÇÕES CIENTÍFICAS EM INFORMÁTICA EM SAÚDE BASEADO EM MODELAGEM AUTOR-TÓPICO

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina, para obtenção do título de Doutor em Ciências

São Paulo
2020

ROBERTO SILVA BAPTISTA

MAPA DE INTERESSES DE AUTORES DE PUBLICAÇÕES CIENTÍFICAS EM INFORMÁTICA EM SAÚDE BASEADO EM MODELAGEM AUTOR-TÓPICO

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina, para obtenção do título de Doutor em Ciências, área de Gestão e Informática em Saúde

Orientador:
Prof. LD. Ivan Torres Pisa

São Paulo
2020

Baptista, Roberto Silva

Mapa de interesses de autores de publicações científicas em informática em saúde baseado em modelagem autor-tópico. / Roberto Silva Baptista. -- São Paulo, 2020.

xix, 106f.

Tese (Doutorado) – Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós-graduação em Gestão e Informática em Saúde.

Título em inglês: Authors´s interests map of health informatics scientific publications based on author-topic modeling.

1. Informática em Saúde, 2. Bibliometria, 3. PubMed, 4. Análise de Redes Sociais, 5. Modelagem de Tópicos.

UNIVERSIDADE FEDERAL DE SÃO PAULO (UNIFESP)
ESCOLA PAULISTA DE MEDICINA (EPM)
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E INFORMÁTICA EM SAÚDE

Coordenadora da Câmara de Pós-graduação e Pesquisa da Escola Paulista de Medicina: Prof. Dra. Monica Levy Andersen, livre docente
Coordenador do Programa: Prof. LD. Ivan Torres Pisa

Roberto Silva Baptista

MAPA DE INTERESSES DE AUTORES DE PUBLICAÇÕES CIENTÍFICAS EM INFORMÁTICA EM SAÚDE BASEADO EM MODELAGEM AUTOR-TÓPICO

Presidente da banca: Prof. LD. Ivan Torres Pisa

BANCA EXAMINADORA

Profa. Dra. Claudia Galindo Novoa
Universidade Federal de São Paulo, Escola Paulista de Medicina

Prof. Dr. Marcelo de Paiva Guimarães
Universidade Federal de São Paulo, Escola Paulista de Medicina

Prof. Dr. Jesús Pascual Mena-Chalco
Universidade Federal do ABC

Profa. Dra. Sueli Mara Soares Pinto Ferreira, livre docente
Universidade de São Paulo, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Dedicatória

Dedico este trabalho à minha esposa Amanda, e
à minha mãe Sonia.

Agradecimentos

Agradeço ao meu orientador Prof. LD. Ivan Torres Pisa pela oportunidade e pelo apoio em todas as fases desta pesquisa.

Agradeço aos amigos e colegas pós-graduandos do Programa que sempre que precisei, contribuíram para o andamento desta pesquisa.

Um agradecimento especial para a minha esposa Amanda que me apoiou não só durante esta pesquisa, mas em toda nossa vida juntos.

Também um agradecimento especial à minha mãe Sonia por entender a minha ausência durante vários momentos deste trabalho.

“A vida começa onde termina sua zona de conforto”.

Neale Donald Walsch

Sumário

Lista de Figuras x
Lista de Tabelas xiii
Lista de Quadros xiv
Lista de Abreviaturas e Símbolos xv
Lista de Publicações xvi
Apoio Financeiro xvii
Resumo xviii
Abstract xix
1 Introdução 20
1.1 Trabalhos Relacionados 23
1.2 Organização do Documento 24
2 Objetivos 25
2.1 Objetivos Específicos 25
3 Materiais e Métodos 26
3.1 Conflitos Éticos e de Interesse 26
3.2 Tipo de Estudo 26
3.3 Materiais 27
3.4 Aspectos Conceituais 27
Pesquisa Quantitativa 28
Autoria e Coautoria 28
Modelagem de Tópicos 29
Análise de Redes Sociais (SNA) 31
3.5 Fluxo de Desenvolvimento da Pesquisa 33
3.6 Etapa 1: Coleta de Artigos do PubMed 34
3.7 Etapa 2: Modelagem Autor-Tópico 37
3.8 Etapa 3: Análise de Redes de Coautorias 40

3.9 Limitações do Estudo 41
4 Resultados 43
4.1 Etapa 1: Coleta de Artigos do PubMed 43
4.2 Etapa 2: Modelagem Autor-Tópico 45
4.2.1 Rotulação dos tópicos obtidos 51
4.2.2 Análise exploratória de cada período de cinco anos 52
4.3 Etapa 3: Análise de Redes de Coautorias 59
4.3.1 Redes de Coautorias 1991-1995 64
4.3.2 Redes de Coautorias 1996-2000 66
4.3.3 Redes de Coautorias 2001-2005 68
4.3.4 Redes de Coautorias 2006-2010 71
4.3.5 Redes de Coautorias 2011-2015 74
5 Discussão 77
5.1 Etapa 1: Coleta de Artigos do PubMed 77
5.2 Modelagem Autor-Tópico 77
5.3 Análise de Redes de Coautorias 79
5.4 Contribuições tecnológicas 80
5.5 Trabalhos Futuros 80
6 Conclusão 82
Referências 83
Anexos 88
Anexo 1 – Aprovação do Comitê de Ética em Pesquisa – UNIFESP – HSP 88
Anexo 2 – PubMed2DB 89
Anexo 3 – TM.NET 92
Anexo 4 – TopicViewer 95
Anexo 5 – Resultados da rotulação de tópicos 100

LISTA DE FIGURAS

Figura 1 – Visão do processo de modelagem de tópicos com LDA. Fonte: Blei, 2012. 29
Figura 2 – Representação gráfica do modelo AT (17).....31
Figura 3 – Exemplos de tipos de rede. À esquerda uma rede não direcionada e à direita uma rede direcionada. Fonte: (40) 32
Figura 4 – Fluxo de desenvolvimento da pesquisa para cumprimento dos objetivos do trabalho. 34
Figura 5 – Passos realizados na etapa 1 de coleta de artigos do PubMed. 35
Figura 6 – Passos realizados na etapa 2 de modelagem autor-tópico. 37
Figura 7 – Passos realizados na etapa 3 de análise de rede de coautorias. 40
Figura 8 – Quantidade de periódicos identificados por ano de publicação. 44
Figura 9 – Quantidade de artigos por ano de publicação. 45
Figura 10 – Recorte da tela de abertura do visualizador do modelo 1991-1995. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico. 46
Figura 11 - Recorte da tela de abertura do visualizador do modelo 1996-2000. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico. 47
Figura 12 - Recorte da tela de abertura do visualizador do modelo 2001-2005. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico. 48
Figura 13 - Recorte da tela de abertura do visualizador do modelo 2006-2010. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico. 49
Figura 14 - Recorte da tela de abertura do visualizador do modelo 2011-2015. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico. 50
Figura 15 – Tópicos 0, 7 e 31 do período de 1991 a 1995.....51
Figura 16 – Tópicos sobre processamento de imagens e sinais médicos. 52
Figura 17 – Lista de artigos do autor James J. Cimino. 53

Figura 18 – Tópico 15 - “informações de saúde na web”. 54
Figura 19 – Tópico 33, rotulado como Telemedicina, as palavras e os autores com maior probabilidade para o tópico. 55
Figura 20 – Recorte de 10 dos 24 artigos no período de 1996-2000 do autor Richard Wooton. 55
Figura 21 – Tópico 46 sobre simulação e treinamento por meio de realidade virtual. 56
Figura 22 – Tópico 18 sobre saúde pública. 56
Figura 23 – Tópico 45 sobre quimioinformática. 57
Figura 24 – Tópico 23 sobre cirurgia assistida por robô 57
Figura 25 – Tópico 41 – “usabilidade de software” 58
Figura 26 – Tópico 33 – “segurança da informação em saúde” 58
Figura 27 – Distribuição de frequências (%) dos tamanhos de caminho. 60
Figura 28 – Distribuição de frequencias (%) de grau. 61
Figura 29 - Representação gráfica das redes obtidas em cada período: (a) 1991 à 1995; (b) 1996 à 2000; (c) 2001 à 2005; (d) 2006 à 2010; (e) 2011 à 2015. 61
Figura 30 – Representação gráfica da rede de 1991-1995 com 7.561 autores e 14.242 coautorias. A figura em alta resolução está disponível em https://bit.ly/2HSveZU 64
Figura 31 – Rede egocêntrica do pesquisador Ove B. Wigertz no período de 1991 a 1995. 65
Figura 32 – Rede egocêntrica do pesquisador J. Robert Beck no período de 1991 a 1995. 66
Figura 33– Representação gráfica da rede de 1996-2000 com 15.782 autores e 35.094 coautorias. A figura em alta resolução está disponível em https://bit.ly/3o9nVMT 67
Figura 34 – Rede egocêntrica do pesquisador Renhold Haux no período de 1996 a 2000. 68
Figura 35 – Representação gráfica da rede de 2001-2005 com 31.257 autores e 78.253 coautorias. A figura em alta resolução está disponível em https://bit.ly/37l8ri5 69
Figura 36 – Rede egocêntrica do pesquisador David W. Bates no período de 2001 a 2005. 70

Figura 37 – Rede egocêntrica do pesquisador J. Marc Overhage no período de 2001 a 2005. 71
Figura 38 – Representação gráfica da rede de 2006-2010 com 55.883 autores e 162.212 coautorias. A figura em alta resolução está disponível em https://bit.ly/33uKYKk 72
Figura 39 - Rede egocêntrica do pesquisador David W. Bates no período de 2006 a 2010. 73
Figura 40 - Rede egocêntrica do pesquisador Xavier Pennec no período de 2006 a 2010. 74
Figura 41 – Representação gráfica da rede de 2011-2015 com 92.355 autores e 342.907 coautorias. A figura em alta resolução está disponível em https://bit.ly/3qpyZrk 75
Figura 42 – Rede egocêntrica do pesquisador David W. Bates e seus coautores no período de 2011 a 2015. 76

LISTA DE TABELAS

Tabela 1 – Número de artigos obtidos por períodos de cinco anos. 45
Tabela 2 - Resultados do pré-processamento.....45
Tabela 3 – Tempos de processamento dos modelos para cada período. 46
Tabela 4 – Número de tópicos identificados por período. 52
Tabela 5 – Métricas globais de caracterização das redes. 59
Tabela 6 – Cinco autores com maior grau em cada período. 62
Tabela 7 – Cinco autores com maior centralidade de intermediação em cada período. 63

LISTA DE QUADROS

Quadro 1 – Dados extraídos para a geração do modelo autor-tópico. 36
Quadro 2 – Lista de periódicos obtida do NLM Catalog. 43
Quadro 3 – Seções do anuário da IMIA em 2015 78

LISTA DE ABREVIATURAS E SÍMBOLOS

CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES	Coordenação de Aperfeiçoamento de Pessoa de Nível Superior
CEP	Comitê de Ética em Pesquisa
DIS	Departamento de Informática em Saúde
UNIFESP	Universidade Federal de São Paulo
NLM	National Library of Medicine
IS	Infomática em Saúde
MeSH	Medical Subject Headings
JCR	Journal Citation Reports
SNA	Análise de Redes Sociais
DBLP	The Computer Science Bibliography
LDA	Latent Dirichlet Allocation
AT	Modelo autor-tópico

LISTA DE PUBLICAÇÕES

Apresentação oral em congresso

Baptista RS, Araújo GD, Teixeira FO, Pisa IT. Scientific Collaboration in Brazilian Health Informatics Scientific Community. In: XXXVI International Sunbelt Social Network Conference Presentation and Poster Abstract [Internet]. California, EUA; p.14–5.

Artigo publicado

Baptista RS, Brito TD de LV, Braun LL, Tenório JM, Pisa IT. Colaboração acadêmica em informática em saúde baseada em análise de redes sociais. J. Health Inform. 7 de dezembro de 2019 Outubro-Dezembro;11(4). 99-6

Artigo submetido

Baptista RS, Pisa IT. 25 years of Health Informatics from an author-topic model perspective. J. Health Inform. Dezembro/2020

APOIO FINANCEIRO

Este projeto de pesquisa recebeu apoio financeiro da CAPES por meio da concessão de bolsa auxílio, nível doutorado, entre agosto de 2016 e julho de 2019.

RESUMO

Baptista RS. Mapa de interesses de autores de publicações científicas em informática em saúde baseado em modelagem autor-tópico [tese – Doutorado]. São Paulo: Departamento de Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo; 2020. 106 f.

Introdução: Informática em saúde (IS) é um campo de pesquisa interdisciplinar em que a quantidade de publicações científicas vem crescendo significativamente. Existem diversas definições sobre IS e pesquisas sobre a estrutura temática da IS. **Objetivo:** O objetivo desta pesquisa é apresentar e validar de forma empírica que uma aplicação de modelagem autor-tópico e análise de redes de coautoria em publicações científicas possibilita evidenciar interesses em pesquisa e estruturas de interação. **Métodos:** Foram coletados artigos de periódicos sobre IS indexados no PubMed e divididos em períodos de 5 anos entre 1991 e 2015. Foi aplicado um modelo autor-tópico para evidenciar interesses de pesquisa dos autores. Também foram aplicadas técnicas de análise de redes sociais para analisar a colaboração entre os autores. **Resultados:** A coleta do PubMed resultou em 69 periódicos e 76.250 artigos. Na modelagem autor-tópico, entre 66% e 84% dos tópicos obtidos foram rotulados. A aplicação de técnicas de análise de redes sociais evidenciou em todos os períodos que os autores com maior centralidade são pesquisadores de extrema relevância em IS. **Conclusão:** Os modelos autor-tópico obtidos apresentaram resultados consistentes, servindo como uma alternativa para evidenciar a evolução da área de IS do ponto de vista dos interesses dos autores identificados pelos tópicos obtidos. A análise de redes de coautoria evidenciou a evolução da estrutura de colaboração global ao longo dos anos, assim como uma visão local da importância dos autores por meio de métricas de centralidade.

Palavras-chave: Mineração de dados (D057225). Informática em saúde (D008490). Análise de redes complexas. Modelagem de Tópicos.

ABSTRACT

Baptista RS. Authors´s interests map of health informatics scientific publications based on author-topic modeling [tese – Doutorado]. São Paulo: Departamento de Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo; 2020. 106 f.

Introduction: Health Informatics (HI) is an interdisciplinary research field which scientific publication is growing significantly. There are several definitions of HI and research on the thematic structure of HI. **Objective:** The objective of this research is to present and empirically validate that an application of author-topic modeling and analysis of co-authorship networks in scientific publications makes it possible to highlight interests in research and interaction structures. **Methods:** Articles from IS journals indexed in PubMed were collected and divided into 5-year periods between 1991 and 2015. An author-topic model was applied to highlight the authors' research interests. Social network analysis techniques were also applied to analyze collaboration between authors. **Results:** The PubMed retrieval resulted in 69 journals and 76,250 articles. In autor-topic modeling, between 66% and 84% of the topics obtained were labeled. The application of social network analysis techniques showed in all periods that the authors with greater centrality are researchers of extreme relevance in IS. **Conclusion:** The author-topic models obtained showed consistent results, serving as an alternative to evidence the evolution of the IS area from the point of view of the authors' interests identified by the topics obtained. The analysis of co-authorship networks showed the evolution of the global collaboration structure over the years, as well as a local view of the importance of the authors through centrality metrics.

Keywords: Data mining (D057225), Health informatics (D008490). Complex network analysis. Topic model.

1 INTRODUÇÃO

Informática em Saúde (IS) é um campo de pesquisa interdisciplinar em que a quantidade de publicações científicas vem crescendo significativamente. Demiris (1) afirma que a IS (neste artigo chamada de Informática Biomédica e em Saúde) é uma área de conhecimento de rápido crescimento e que depende da ativa colaboração entre diferentes disciplinas e profissões.

Embora existam diferentes definições, a maioria se refere à IS com referência às ciências da saúde e o apoio das tecnologias da informação e comunicação (2). Segundo Greenes e Siegel (3), esta característica interdisciplinar traz desafios especiais inclusive para a National Library of Medicine (NLM). Neste trabalho os autores abordaram como delimitar as fronteiras de disciplinas que compõem um campo interdisciplinar como a IS. Uma das conclusões foi que medidas puramente bibliométricas como o fator de impacto não refletem a importância de uma publicação em IS. Nesse sentido para tentar entender melhor a estrutura interdisciplinar da IS, Morris e McCain (2) realizaram uma análise de citações e co-citações entre artigos científicos. Como um de seus resultados os autores evidenciaram a interdisciplinaridade da IS por meio de análise fatorial dos dados de co-citação onde vários periódicos fazem uso de múltiplos fatores ou disciplinas.

No trabalho de DeShazo e colaboradores (4) foi utilizado o MeSH para caracterizar a tendência de publicações em informática em saúde. Os autores coletaram do PubMed os artigos que possuíam o termo “Medical Informatics” como [MeSH Terms] ou [Major]. Também foi utilizado o JCR Category “medical informatics” para os periódicos encontrados na coleta. Os autores identificaram um crescimento exponencial de artigos em IS no período entre 1987 e 2006 assim como um crescimento médio anual de artigos em IS de 12% ao ano. Este crescimento é maior que o crescimento anual total do PubMed (4%) e maior que algumas áreas como Medicina (5%) e Saúde Pública (8%). Foi concluído pelos autores que esse crescimento teve três dimensões primárias: O número de periódicos de Informática em Saúde, o número de artigos de Informática em Saúde indexados e o aumento do número de artigos de IS publicados em periódicos que não são sobre IS.

Lyu e colaboradores (5) analisaram os termos MeSH emergentes encontrados nos artigos coletados do PubMed entre 2000 e 2011. Foram coletados os artigos no

período que possuíam “Medical Informatics” com [Major]. Para cada artigo, foram analisados os seus termos MeSH e calculadas as taxas de incremento para cada termo, sendo considerados termos emergentes aqueles que possuíam maior taxa de incremento. Em seguida foi calculado o fator de perspectiva (perspective factor, PF) para cada periódico. O PF indica a média de utilização de termos emergentes em cada artigo do periódico em estudo.

Outra abordagem é o estudo das coautorias em artigos científicos, consideradas evidências confiáveis de colaboração científica (6). A colaboração científica é geralmente representada em redes sob a ótica de análise de redes sociais (*social network analysis,* SNA).

SNA é um tema interdisciplinar que abrange diferentes áreas do conhecimento como sociologia, matemática, física e estatística (7). Muitos sistemas podem ser organizados em rede, conjunto de nós conectados em pares por arestas ou ligações. Exemplos desses sistemas são as redes sociais, a internet e as redes de citações entre trabalhos científicos. Em redes sociais, pessoas são representadas por nós e os relacionamentos entre as pessoas são representados por arestas (8). Nesta abordagem o relacionamento entre pessoas é o foco principal de estudo e as características individuais são tratadas como secundárias (9).

Em redes de colaboração científica baseadas em coautorias de artigos científicos, pesquisadores são considerados entidades sociais e as coautorias entre os pesquisadores são consideradas ligações (10). Estudos de redes de colaboração científica podem contribuir para elaboração de indicadores alternativos para agências de fomento à pesquisa. Em 2010 Freire e Figueiredo apresentaram uma métrica de ranqueamento de pesquisadores brasileiros em ciência da computação (11). Esta métrica buscou representar a intensidade da colaboração científica por meio das coautorias em publicações de Ciência da Computação indexadas na base DBLP. A métrica apresentada foi comparada com a avaliação subjetiva realizada pelo CNPq para concessão de bolsas de produtividade. Os resultados indicaram que a métrica foi eficaz para identificar a influência dos pesquisadores.

Em ciências da saúde, mais especificamente na área de psiquiatria, Wu e Duan (12) analisaram a colaboração científica entre instituições e entre países por meio de 36557 artigos sobre psiquiatria entre os anos de 2003 e 2012. Os resutados

indicaram um crescimento na colaboração científica tanto entre pesquisadores quanto entre instituições e entre países.

Em pesquisa translacional, Vacca e colaboradores (13) investigaram a sugestão de colaboração entre pesquisadores de diferentes áreas que não haviam trabalhado juntos anteriormente. Os resultados indicaram serem viáveis tais sugestões para a montagem de grupos de trabalho em pesquisa translacional.

Em recuperação da informação (*Information Retrieval*, IR) métodos de modelagem de tópicos (*Topic Models*, TM) vêm sendo cada vez mais estudados (14, 15, 16). Blei e colaboradores, 2003 definem TM como uma modelagem probabilística para descobrir a estrutura semântica de um conjunto de documentos baseada em padrões de uso de palavras e ligando documentos com padrões similares. TM vem sendo aplicada a vários tipos de documentos como emails, textos de jornais e resumos de artigos científicos. Um dos avanços em TM foi o método de modelagem autor-tópico (*author-topic model, AT)* proposto por Rosen-Zvi e colaboradores (17). Neste, os autores afirmam que incluindo a autoria dos documentos, é possível por exemplo mapear os interesses acadêmicos de pesquisadores. Neste método, baseado no método clássico de TM, LDA (*latent dirichlet allocation*) (14), dado um conjunto de documentos, cada autor é associado a uma distribuição de frequência de tópicos e cada tópico associado a uma distribuição de frequência de palavras extraídas do conjunto de documentos. Assim um dado documento é modelado pela distribuição de frequência de tópicos de cada co-autor deste documento.

A tese aqui apresentada trata da aplicação de técnicas de modelagem autor-tópico e análise de redes de coautoria em um conjunto de artigos de periódicos indexados no PubMed sobre informática em saúde. O objetivo foi evidenciar interesses de pesquisa de autores e também evidenciar a colaboração entre os mesmos.

Esta pesquisa faz parte do grupo de pesquisa Saúde 360º (saude360.unifesp.br) que tem foco em descoberta de conhecimento e mineração de dados em saúde.

Uma das pesquisas realizadas no grupo foi o de Colepícolo em 2018 (18) denominada “Epistemologia da Informática em Saúde: teoria e prática”. Por meio de seus resultados foi possível inferir que a área de informática em saúde reúne dois tipos de conhecimentos. O primeiro sendo a base específica, composta por conhecimento oriundos de outras áreas, especificamente: ciências comportamentais, ciên-

cias naturais, ciências biológicas e ciência da informação. O segundo tipo é o corpo do conhecimento, obtido da própria área: informática biomédica, informática médica, informática em enfermagem, informática odontológica e informática aplicada às outras ciências da saúde não especificadas.

Teixeira em 2011 (19) em sua dissertação de mestrado denominada “Classificação e indexação de artigos científicos internacionais de informática em saúde” buscou realizar uma classificação e indexação de um conjunto de artigos extraídos do ISI Web of Knowledge, com a utilização do Journal Decriptor Index (JDI) (20).

Em relação a análise de redes de coautoria, Costa e colaboradores (21) realizaram um estudo para mapear as conexões entre estudantes e profissionais da área de informática em saúde. Neste trabalho, com informações obtidas a partir dos currículos da Plataforma Lattes CNPq, foi proposta uma medida de popularidade denominada Índice Pisa de Popularidade (IPP).

Em 2012, iniciei meu interesse sobre o tema de análise de redes de coautoria, quando iniciei um trabalho de ampliação da base utilizada por Costa e colaboradores (21) que utilizou técnicas e métricas clássicas de análise de redes sociais para avaliar a colaboração em IS. Resultados preliminares foram apresentados na primeira edição da Conferência Européia sobre Análise de Redes em 2014 (22). Por ocasião da conferência tive contato com pesquisas ligadas ao tema de modelagem de tópicos e iniciei meus estudos acerca do tema. Posteriormente em 2016 novos resultados foram apresentados na 36ª edição da International Sunbelt Social Network Conference (23).

Outros participantes do grupo também demonstraram interesse nas técnicas de análise de redes sociais e de modelagem de tópicos. Como exemplo em 2016, Brito aplicou técnicas de análise de redes sociais em sua dissertação de mestrado para analisar a colaboração nos grupos de interesse especial (SIG) da Rede Universitária de Telemedicina (RUTE) (24).

1.1 Trabalhos Relacionados

Redes de coautoria é um tema de grande interesse de pesquisadores e que contribui para compreensão dos padrões de colaboração científica entre pesquisadores. Uma das grandes contribuições foi o estudo de Newman (10). Sobre informá-

tica em saúde, Jeong e colaboradores (25) analisaram a estrutura de conhecimento da área de informática em saúde na Coréia do Sul por meio da aplicação de técnicas de análise de redes sociais. Também sobre informática em Saúde, Baptista e colaboradores (26) analisaram a rede de colaboração em informática em saúde no Brasil por meio de aplicação de técnicas de análise de redes sociais.

Chen e colaboradores (27), aplicaram a modelagem de tópicos em títulos e resumos de artigos que continham a palavra ‘ginseng’ do PubMed para a descoberta de tópicos e sua evolução. Os autores identificaram um tópico relacionado ao uso do ginseng como suplemento dietético bem como outro tópicorelacionado ao plantio do ginseng. Wang e colaboradores utlizaram a modelagem de tópicos baseado no LDA (*Latent Dirichlet Allocation*) e concluíram que sua aplicação além de ser uma alternativa para reconhecer fatos já conhecidos, também auxilia na descoberta de novos tópicos relevantes (28).

Kongthon e colaboradores (29) aplicaram a modelagem autor-tópico para melhoria do processo de revisão da literatura tradicional.

Outra aplicação da modelagem autor-tópico é a seleção de revisores de artigos em periódicos ou conferências. Kusumawardani e Khairunnisa (30) avaliaram a utilização da modelagem autor-tópico para a atribuição de revisores à artigos científicos submetidos. Os resultados preliminares indicaram que os revisores propostos possuíam os conhecimentos necessários para revisarem os artigos submetidos.

1.2 Organização do Documento

O presente documento está organizado nos seguintes capítulos:

- Capítulo 1: o capítulo corrente, contendo a contextualização do trabalho e a descrição dos objetivos, geral e específicos;
- Capítulo 2: Trabalhos relacionados e aspectos conceituais;
- Capítulo 3: materiais e métodos utilizados para desenvolver a pesquisa e atingir os objetivos;
- Capítulo 4: resultados alcançados pela pesquisa;
- Capítulo 5: discussão sobre os resultados relatados no Capítulo 4;
- Capítulo 5: conclusões da pesquisa realizada;

2 OBJETIVOS

O objetivo principal desta pesquisa é apresentar e validar de forma empírica que uma aplicação de modelagem autor-tópico e análise de redes de coautoria em publicações científicas possibilitam evidenciar interesses em pesquisa e estruturas de interação.

Esta pesquisa buscou possibilitar responder perguntas como:

- Um modelo autor-tópico pode ser uma alternativa de caracterização dos interesses de autores de publicações científicas da área de informática em saúde?
- Que autores são mais populares em cada tópico?
- Que autores colaboram mais em um determinado tópico?
- Quais pesquisadores têm interesses acadêmicos semelhantes e quais já escreveram artigos juntos? E quais destes não escreveram juntos?

2.1 Objetivos Específicos

Os objetivos específicos desta pesquisa são:

Objetivo 1: Construir e analisar modelo autor-tópico (AT) por períodos de cinco anos compreendidos entre os anos de 1991 a 2015, para categorizar interesses acadêmicos em informática em saúde baseado em resumos e autorias de artigos científicos.

Objetivo 2: Construir e analisar rede de coautoria por períodos de cinco anos compreendidos entre os anos de 1991 a 2015.

3 MATERIAIS E MÉTODOS

Neste capítulo estão descritos os materiais que foram utilizados para o desenvolvimento da pesquisa e os métodos utilizados para cumprir o objetivo principal e específicos do trabalho.

3.1 Conflitos Éticos e de Interesse

Este estudo foi realizado junto ao grupo de pesquisa Saúde 360º (https://saude360.unifesp.br), associado ao Programa de Pós-graduação em Gestão e Informática em Saúde, Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (UNIFESP). Foi aprovado pelo Comitê de Ética em Pesquisa (CEP) sob o número 0665/2015 (Anexo 1 – Aprovação do Comitê de Ética em Pesquisa – UNIFESP – HSP, pág. 88).

Cabe salientar que todos os dados utilizados neste estudo são dados abertos e públicos disponíveis na plataforma web do PubMed. Os nomes dos autores aqui identificados constam de coautorias de artigos científicos, que após sua publicação em periódico, se tornam públicos.

A despeito do pesquisador e seu orientador atuarem academicamente na área de informática em saúde, ambos declaram não existir qualquer conflito de interesse quanto aos resultados dos grupos de autores, tópicos e suas associações. Nenhuma manipulação ou privilégio foi gerado nos modelos matemáticos aplicados para que determinados autores, grupos, tópicos, periódicos ou instituições ocorressem em maior ou menor ordem nos resultados apresentados.

3.2 Tipo de Estudo

A proposta deste estudo foi realizar uma pesquisa quantitativa composta por uma análise correlacional no que tange à modelagem autor-tópico para categorização dos interesses acadêmicos de pesquisadores em informática em saúde. Também faz parte do estudo a realização de uma análise de rede de coautoria de artigos científicos da mesma área.

3.3 Materiais

A infraestrutura de hardware utilizada para realizar a pesquisa e desenvolver o trabalho foi:

- Microcomputador Notebook Dell 7000 Series, processador Intel® I7® 5500U (2.40 GHz, 64-bit), 8 GB de memória RAM, 240 GB de SSD, para pesquisa bibliográfica, construção e testes dos modelos, análise dos resultados e redação da tese.
- Servidor web Intel® CPU, 3 GHz com 4 núcleos, 16 GB de memória RAM e 2 TB de HD, para hospedagem da aplicação web TopicViewer.

Adicionalmente os seguintes softwares e tecnologias foram utilizados para dar suporte ao desenvolvimento da pesquisa:

- Ambiente de desenvolvimento integrado Microsoft Visual Studio 2012 Express Edition (Microsoft, Inc) de uso gratuito.
- Linguagem de programação C# com framework .NET® versão 4.5.1 (http://www.microsoft.com/net).
- Sistema Gerenciador de Banco de Dados (SGBD) SQL Server® Express versão 2012 (http://www.microsoft.com/express).
- Pacote Estatístico R versão 2.15 (http://www.r-project.com).
- Pacotes adicionais para SNA para o R: igraph, sna, network, statnet.
- Pacote adicional para modelagem de tópicos para o R: topicmodel.
- Pacote XML, RCurl, plyr, reports, stringr e RISmed para acesso e extração de dados do PubMed.
- Softwares para modelagem e visualização de SNA Gephi versão 0.8 (http://gephi.org/) e Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

3.4 Aspectos Conceituais

Esta pesquisa baseia-se na realização de uma pesquisa quantitativa tendo em sua essência a aplicação das técnicas de modelagem autor-tópico e de análise

de redes sociais. Seguem apresentados os principais conceitos subjacentes à condução metodológica do estudo.

Pesquisa Quantitativa

Na pesquisa empírica, existem duas abordagens que podem ser consideradas, quantitativa e qualitativa. Uma das definições da abordagem quantitativa(31) é associada a coleta e conversão de dados em formato numérico para realizar cálculos estatísticos que resultarão em conclusões generalizadas ou específicas para validação de hipóteses. Segundo Wainer(32) a pesquisa quantitativa se baseia numa visão positivista em que as variáveis observadas são objetivas, ou seja, diferentes observadores obterão os mesmos resultados, não há desacordo do que é melhor ou pior para os valores dessas variáveis objetivas e medições numéricas são consideradas mais ricas que descrições verbais.

Autoria e Coautoria

Garcia e colaboradores (33) comentam que uma das definições de autoria é o direito legal do autor sobre seu texto. E acrescentam que além disse é um diretio moral e econômico sobre seu texto. Os autores lembram que para Foucault, o autor é um organizador do conhecimento, que cria um significado que busca trazer relevância e credibilidade. No meio científico, reconhecer a autoria científica possibilita avaliar a produtividade científica de um pesquisador.

A coautoria científica pressupõe a participação ativa na produção de conhecimento científico, e é caracterizada pela coparticipação na redação total ou parcial de uma pesquisa (34). A coautoria científica é um resultado factual da colaboração científica entre pesquisadores.

Uma das maneiras de se medir a colaboração entre pesquisadores é o coeficiente de colaboração (CC), que indica o número médio de autores por artigo (35). Existem variações desta medida, mas o conceito em si é semelhante.

Garcia e colaboradores (33) alertam que como a produção científica de um autor está relacionada a diferentes aspectos de ganhos de status e fomento à pesquisa, questões éticas são muito discutidas sobre o tema. Nesse sentido a COPE (Committee on Publication Ethics) (36), que foi criada para discutir, tratar e educar

as partes envolvidas em publicações científicas sobre questões éticas da comunicação científica, possui uma iniciativa dedicada a ética da autoria e da coautoria. Outras iniciativas passam por conflitos de interesse, propriedade intelectual e dados e reprodutibilidade.

Modelagem de Tópicos

Segundo Blei e colaboradores (14) a modelagem probabilística de tópicos, ou somente modelagem de tópicos, representa uma abordagem que visa evidenciar temas ou tópicos em grandes conjuntos de documentos. Nesse trabalho é apresentado o Latent Dirichlet Allocation (LDA) que representou um grande marco no estudo de modelagem de tópicos e se tornando muito popular. O LDA (14) é constituído por um modelo bayesiano hierárquico de três camadas em que cada documento é modelado por uma distribuição de probabilidade de tópicos e cada tópico é modelado por uma distribuição de probabilidades de palavras. Na Figura 1 é apresentada uma visão do processo de modelagem de tópicos com LDA.

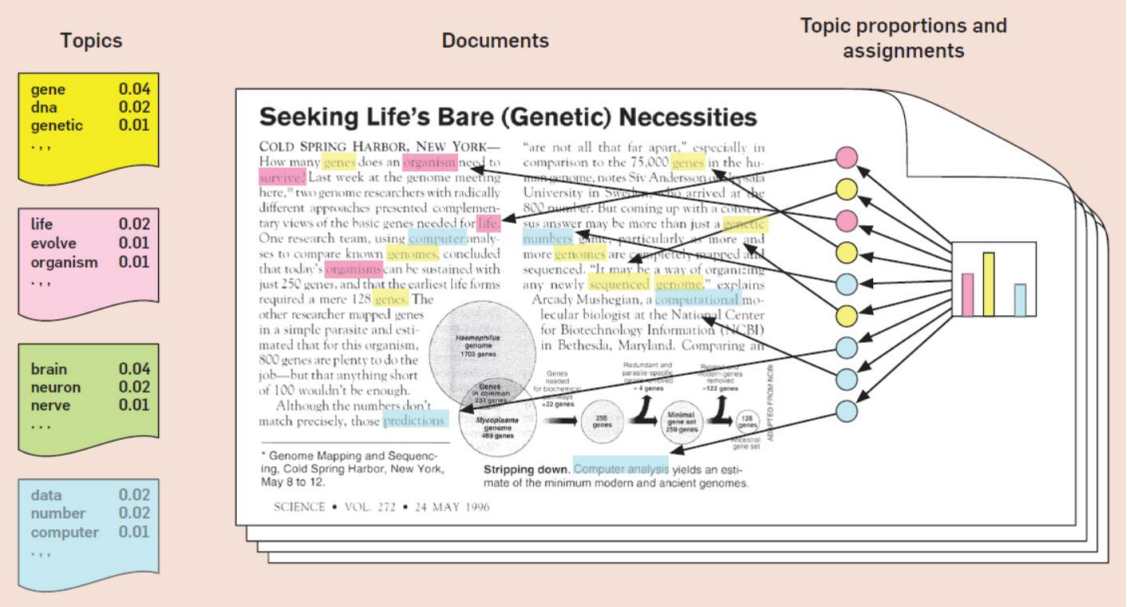


Figura 1 – Visão do processo de modelagem de tópicos com LDA. Fonte: Blei, 2012.

Diversos autores propuseram derivações e complementações utilizando o LDA como base (16,17, 27)

Em 2007 Blei e Lafferty (16) propuseram a modelagem de tópicos correlacionados, que, ao levar em consideração a correlação entre tópicos, por exemplo, um

documento sobre geologia é mais provável de ser sobre arqueologia do que genética.

O LDA foi criado como sendo uma técnica de aprendizado de máquina não supervisionado, ou seja, não há uma variável alvo a priori, categoria ou rótulo para que o conjunto de dados seja agrupado. Ao invés disso, os possíveis grupos emergem do próprio conjunto de dados que foi submetido. Uma das variações do LDA foi o Supervised Topic Model (38) que considera um rótulo prévio para os documentos, por exemplo a quantidade de usuários que classificaram um artigo como interessante numa comunidade on-line.

Neste estudo foi utilizada a modelagem autor-tópico (AT) (17). Neste método, que tem como base o LDA, a principal diferenciação é que cada documento é representado pelo conjunto de distribuições de probabilidade de tópicos de cada autor que participou da coautoria do documento e não somente uma única distribuição de probabilidade de tópicos.

No modelo AT, assim como no LDA, precisamos informar os hiper-parâmetros de inicialização do modelo, alfa (α) e beta (β). No LDA, α é uma matriz que representa a densidade de tópicos para cada documento e β é uma matriz que representa a densidade do vocabulário para cada tópico. Cabe salientar que há duas maneiras de se escolher α e β. Um maneira é quando se tem conhecimento prévio sobre as densidades, neste caso são informadas matrizes em que seus elementos variam conforme o conhecimento prévio. A segunda é quando não se tem conhecimento prévio sobre as densidades, neste caso é definido um mesmo valor para todos os elementos de cada matriz. No modelo AT a diferença é que o hiper-parâmetro α é uma matriz que representa a densidade de tópicos para cada autor, e não documento. A Figura 2 apresenta uma representação gráfica do modelo AT. Nesta figura, **D** representa o conjunto de documentos, **A** representa o conjunto de autores e suas distribuições de tópicos, **T** representa o conjunto de tópicos e suas distribuições de palavras. $\mathbf{a_d}$ representa o conjunto de autores de um documento **d**. $\mathbf{N_d}$ representa o conjunto de palavras de um documento **d**. **w** é a palavra obtida de um tópico **z** que foi escolhido da distribuiçãoo de tópicos do autor **x** pertencente ao conjunto $\mathbf{a_d}$.

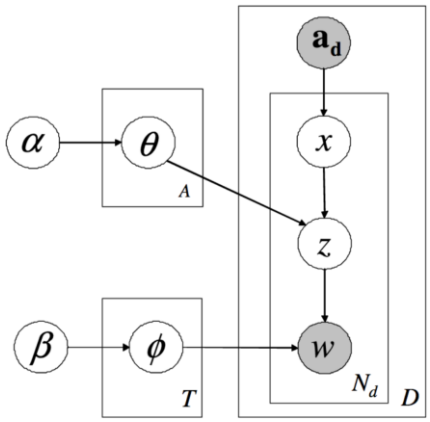


Figura 2 – Representação gráfica do modelo AT (17)

Análise de Redes Sociais (SNA)

A chamada análise de redes sociais (*social network analysis*, SNA) pode ser definida como um conjunto de estratégias para investigar estruturas sociais (9) em que o estudo das relações entre os entes sociais é privilegiado em relação às características individuais de cada ente social. Redes sociais são redes formadas por indivíduos que se conectam, e interagem, seja em função de ligação familiar, amizade, trabalho entre outros. Para representar e estudar tais relações, a SNA buscou se apoiar na teoria dos grafos, subárea da matemática desenvolvida por Leonhard Euler. A análise de redes sociais também é classificada como um exemplo de análise de redes complexas, assim como a análise de redes de relações de conceitos em um dicionário ou tesauro e redes de citações de trabalhos científicos (39).

Uma rede ou Grafo é definido como um conjunto de nós (ou vértices) e um conjunto de ligações (ou arestas) entre estes. Se as ligações forem simétricas entre os nós a rede é considerada não direcionada. Se as ligações forem assimétricas a rede é considerada direcionada. Como exemplo, uma rede de co-citação de artigos é direcional, já uma rede de coautoria de artigos é uma rede não direcionada. Na Figura 3 são apresentados exemplos dos dois tipos de redes no que se refere ao tipo de ligação.

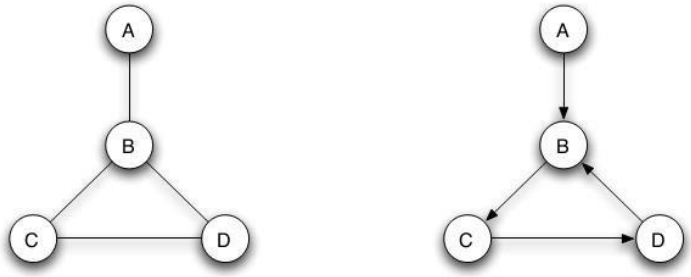


Figura 3 – Exemplos de tipos de rede. À esquerda uma rede não direcionada e à direita uma rede direcionada. Fonte: (40)

Uma série de propriedades podem ser obtidas a fim de caracterizar uma rede como por exemplo:

- Número de nós: a quantidade de nós de uma rede. Também conhecido como tamanho da rede;
- Número de ligações: a quantidade de ligações entre os nós de uma rede;
- Densidade da rede: A densidade de uma rede é medida pela razão entre o número de ligações existentes e o número máximo de ligações possíveis de uma rede com o mesmo número de nós. A densidade varia entre 0 e 1, sendo 1 uma rede totalmente conectada, onde cada nó possui uma ligação com qualquer outro nó da rede.
- Grau médio: Considerando que o grau de um nó é a quantidade de ligações que este possui, o grau médio de uma rede é a média entre os graus de todos os nós da rede. Cabe destacar que em se tratando de uma rede direcionada, existem três tipos de grau: grau de entrada, grau de saída e grau total, que é a soma dos graus de entrada e saída. No caso de uma rede não direcionada, existe somente um tipo de grau pois a ligação não possui direção.
- Caminho médio: Um caminho entre dois nós quaisquer de uma rede é representado pela menor quantidade de ligações que existem entre eles. O caminho médio é a média de todos os caminhos(menores) entre todos os pares de nós de uma rede.
- Diâmetro da rede: é maior caminho encontrado entre todos os pares de nós de uma rede. O diâmetro de uma rede pode variar entre 1 e a quantidade de nós menos 1 (n-1);

- Componente gigante: uma rede é composta de subredes, ou componentes, que são um subconjunto de nós que possuem pelo menos um caminho entre todos os pares de nós. Componente gigante é o componente que possui o maior número de nós numa rede;

Além das métricas globais para caracterização de redes, métricas de rede locais de centralidade são métricas para caracterização individual dos nós que indicam importância, popularidade, entre outros. Alguns exemplos são:

- Grau: a quantidade de nós que um nó possui ligações;
- Grau ponderado: a quantidade de ligações que um nó possui;
- Centralidade de intermediação: a quantidade de caminhos entre pares de nós que um determinado nó faz parte.

Para este trabalho foi considerado o estudo de redes de coautoria de artigos científicos, um exemplo de rede social, em que dois pesquisadores possuem uma ligação se ambos são coautores de um mesmo artigo científico. Um artigo científico com dois ou mais autores indica que estes se conhecem e colaboraram entre si (10).

3.5 Fluxo de Desenvolvimento da Pesquisa

O fluxo de desenvolvimento deste estudo foi realizado em três grandes etapas para atingir os objetivos propostos. A primeira etapa resume-se à coleta dos resumos e coautorias de artigos científicos. A segunda etapa representa o pré-processamento e posterior modelagem autor-tópico e a terceira etapa representa a análise de rede de coautorias (Figura 4).

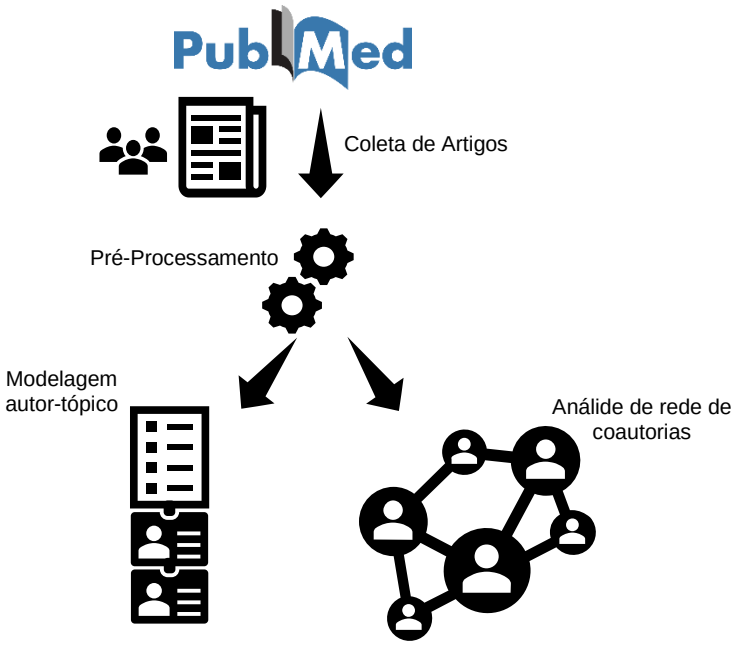


Figura 4 – Fluxo de desenvolvimento da pesquisa para cumprimento dos objetivos do trabalho.

A seguir estão apresentados os detalhamentos de cada grande etapa realizada nesse estudo para atingimento dos objetivos.

3.6 Etapa 1: Coleta de Artigos do PubMed

O PubMed (http://www.ncbi.nlm.nih.gov/pubmed) foi utilizado como fonte de dados neste trabalho. Sua escolha se deve por ser a principal ferramenta internacional de busca e recuperação de informação em publicações científicas em ciências da saúde. Por meio do PubMed pode-se ter acesso a mais de 30 milhões de registros de publicações científicas. Destas publicações, 13 milhões são providas de seus respectivos resumos (41).

O PubMed agrupa registros de publicações em três diferentes fontes da National Library of Medicine (NLM). São estas: Medline, PubMed Central (PMC) e Bookshelf. O Medline compreende a maior parte dos registros de publicações indexadas do PubMed e possui principalmente registros oriundos de periódicos selecionados segundo seus critérios. O PMC é o segundo maior componente do PubMed e possui publicações na íntegra selecionados segundo seus critérios e políticas. O Bookshelf é uma fonte livros e capítulos individuais de livros, relatórios, bases de dados e outros documentos na íntegra.

Esta etapa foi dividida em três passos. O primeiro passo foi a seleção dos periódicos indexados no catálogo da National Library of Medicine (NLM Catalog) que possuíam a área de informática em saúde em seu escopo. O segundo passo foi a coleta de todos os artigos de cada periódico selecionado e o terceiro passo foi a extração dos resumos, incluíndo título do periódico, ano de publicação, título do artigo e resumo do artigo e as respectivas coautorias. Esse conjunto de dados foi dividido em períodos de cinco anos.

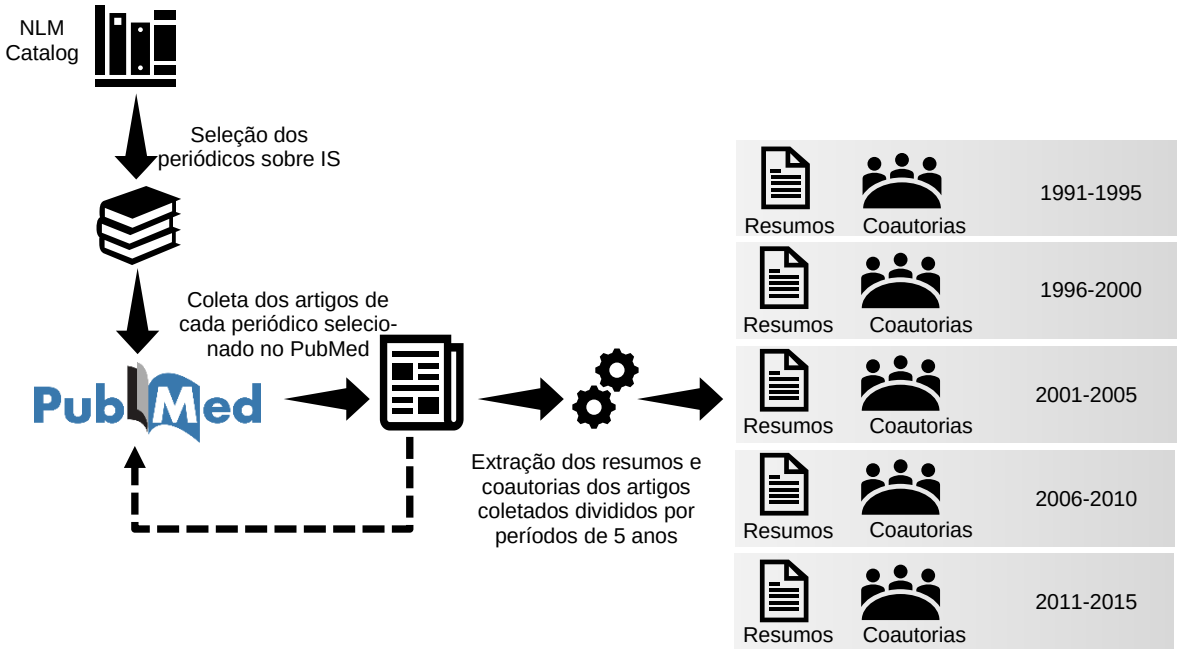


Figura 5 – Passos realizados na etapa 1 de coleta de artigos do PubMed.

Para esta etapa foi utilizado o software R em conjunto com o pacote RISmed (http://cran.r-project.org/web/packages/RISmed) para a busca e coleta dos títulos de cada periódico (primeiro passo) e seus respectivos artigos (segundo passo). O pacote RISmed implementa todos os acessos (via padrão API) aos serviços disponibilizados no Entrez Programming Utilities (42), que dentre outras funções possibilita a busca e coleta de metadados de artigos do PubMed.

No primeiro passo, para selecionar os periódicos indexados pelo PubMed que possuíam informática em saúde em seu escopo foi definida e submetida a seguinte estratégia de busca na base NLM Catalog:

```
('"medical informatics"[MeSH Terms] OR ("medical"[All
Fields] AND "informatics"[All Fields]) OR "medical infor-
matics"[All Fields] OR ("health"[All Fields] AND "infor-
matics"[All Fields]) OR "health informatics"[All Fields]
OR ("nursing"[All Fields] AND "informatics"[All Fields])
OR "nursing informatics"[All Fields] OR ("dental"[All
Fields] AND "informatics"[All Fields]) OR "dental infor-
matics"[All Fields]) AND currentlyindexed[All])'
```

Cabe salientar que o último termo desta estratégia de busca – `currentlyindexed[All]` – tem o objetivo de limitar a busca somente para periódicos que estão indexados no Medline em sua versão corrente (43).

No segundo passo foram coletados todos os artigos de cada periódico obtido no passo anterior. Para cada periódico uma estratégia de busca foi submetida com a seguinte construção: "[Journal]=<título-do-periódico>", no qual <título-do-periódico> se refere ao título abreviado de cada periódico obtido no primeiro passo.

No terceiro passo toda a estrutura de arquivos coletada foi transformada e armazenada em banco de dados relacional para facilitar a extração e preparação dos arquivos de resumos e coautorias. A descrição deste método de extração é apresentada no Anexo 2 (pág. 89). Como critérios de exclusão foram descartados artigos com data de publicação anterior ao ano de 1991 e artigos com data de publicação superior a 2015. Também foram descartados artigos que não possuíam resumos disponíveis no PubMed. Como saída deste passo foi obtido um par de arquivos texto para cada período de cinco anos. O primeiro arquivo chamado `resumos.csv` representa o conjunto de artigos em que cada linha representa um artigo. Cada linha possui como conteúdo o título do artigo e seu resumo separado por um caractere de espaço " ". O segundo arquivo chamado `coautorias.csv` representa o conjunto de coautorias dos respectivos artigos representados no arquivo `resumos.csv`. Cada linha possui a relação de coautoria de cada artigo, estes são separados por um caractere de ponto e vírgula ";". Cabe salientar que a ordem das linhas deve ser a mesma nos dois arquivos. Como exemplo, o artigo a que se refere a linha 1, no arquivo `resumos.csv` se refere a coautoria encontrada na linha 1 do arquivo `coautorias.csv`. Os campos utilizados da estrutura de registros de artigo do Pubmed são apresentados no Quadro 1. Cabe salientar que os campos "ano de publicação" e "título do periódico" não foram utilizados como dados de entrada para a modelagaem autor-tópico. O campo "ano de publicação" somente foi usado para identificação dos períodos e o campo "título do periódico" somente foi usado após a modelagem autor-tópico para manter a informação do periódico em que o artigo foi publicado.

Quadro 1 – Dados extraídos para a geração do modelo autor-tópico.

Campo Obtido	**Campo Origem**
Título do artigo	ArticleTitle.articleTitle
Título do periódico	Journal.ISOAbbreviation
Ano de publicação	PubmedPubDate.year

Resumo	AbstractText.value
Lista de autores	Author.LastName + Author.ForeName OR Author.CollectiveName

3.7 Etapa 2: Modelagem Autor-Tópico

Nesta etapa estão descritos os passos para a modelagem autor-tópico utilizada nesta pesquisa.

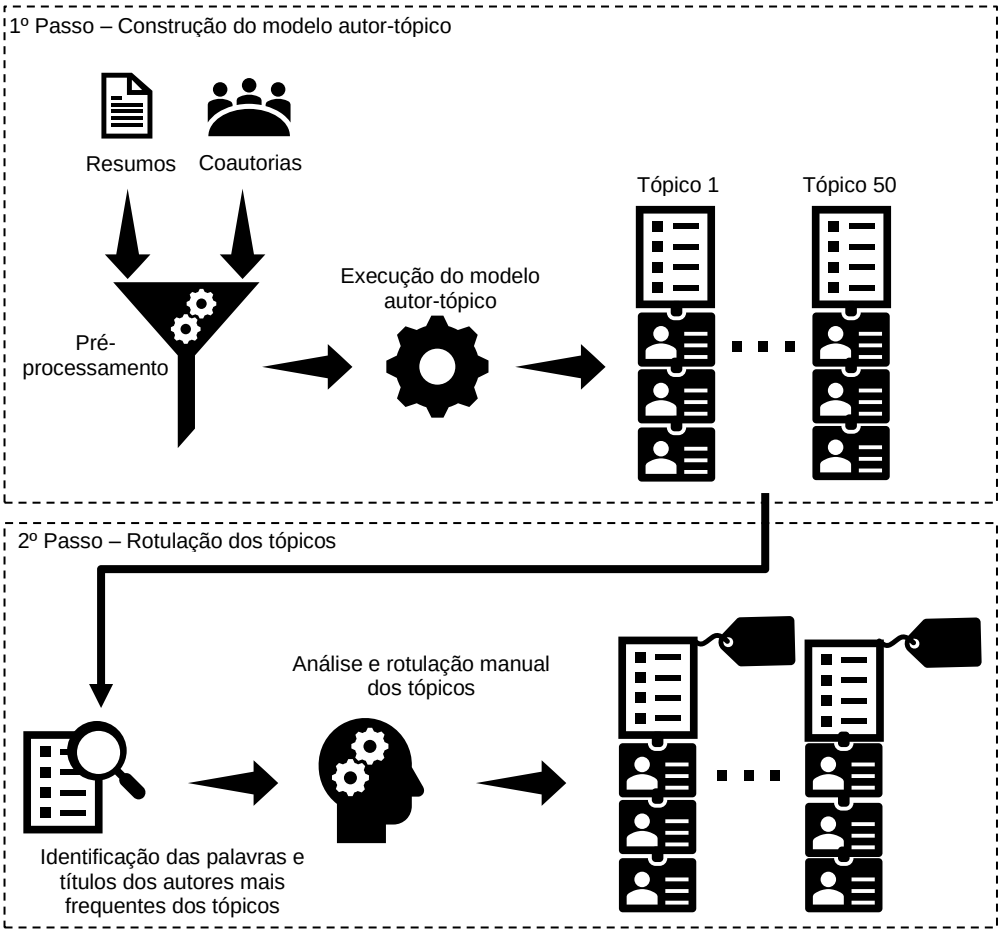


Figura 6 – Passos realizados na etapa 2 de modelagem autor-tópico.

Para a aplicação da modelagem autor-tópico nesta pesquisa foi necessário o desenvolvimento pelo próprio pesquisador de um software para sua implementação e execução porque até o momento desta etapa da pesquisa não foi encontrada na literatura uma implementação disponível para utilização. Uma descrição desta implementação, chamada de TM.NET, pode ser encontrada no Anexo 3 (pág. 92).

O primeiro passo desta etapa inicia-se com o pré-processamento. O pré-processamento é uma etapa extremamente importante em diferentes abordagens como mineração de textos (*text mining*) e processamento de linguagem natural (*natural language processing*, NLP). Alguns exemplos de técnicas que podem ser usadas são: remoção de palavras extremamente frequentes no corpus, radicalização de palavras (*stemming*), geração de n-gramas (por exemplo bigramas, trigramas). Aqui neste passo foram aplicadas as seguintes técnicas:

- Conversão para minúsculo: transforma todos os caracteres alfabéticos encontrados em caracteres minúsculos para evitar que as rotinas computacionais posteriores não considerem como palavras distintas, palavras idênticas que apenas possuem diferenças entre maiusulas e minúsculas.
- Remoção de números e caracteres não alfabéticos.
- Remoção de *stopwords*: Remoção do corpus de palavras muito frequentes encontradas numa lista em um arquivo de texto. Para esta pesquisa foi utilizada a lista obtida pela função *stopwords("smart")* do pacote TM (44) disponível no software R. Foram acrescentadas manualmente pelo próprio pesquisador palavras comuns ao domínio das publicações científicas: *introduction, method, discussion, conclusion*, *study*, *research*, *paper*, entre outras.
- Remoção de espaços extras.
- Remoção de palavras com alta e baixa frequência no corpus. Foram removidas palavras que ocorrem em menos de 0,1% dos resumos ou acima de 99% dos resumos.
- Geração de n-gramas, mais especificamente bigramas e trigramas.

Ao final do pré-processamento um conjunto de dados para cada período de cinco anos é retornado para posterior execução do algoritmo autor-tópico.

Para cada período de cinco anos os respectivos conjuntos de dados obtidos no pré-processamento foram submetidos ao gerador de modelo autor-tópico do software TM.NET para geração dos modelos autor-tópico. Como parâmetros de entrada foram utilizados k=50 (número de tópicos) e os hiperparâmetros alfa=0,1 e beta=0,01. Estes valores de alfa e beta foram utilizados para todos os elementos das matrizes alfa e beta, pois não havia conhecimento sobre a densidade da distribuição

de palavras nos tópicos e nem sobre a densidade de tópicos para os autores. Para cada período foi utilizado o amostrador de gibbs com a mesma configuração, descarte das primeiras 1.000 iterações e execução de 2.000 iterações.

Assim para cada período de cinco anos foram obtidos:

- uma matriz de distribuição de frequência de palavras por tópico (phi);
- uma matriz de distribuição de frequência de tópicos por autor (theta).

Também foi obtida uma matriz de similaridade por cosseno entre autores baseada nas distribuições de frequência dos tópicos dos autores.

No passo 2, para apoiar a análise do conteúdo obtido foi desenvolvida uma aplicação web para visualização dos tópicos identificados em cada período chamada TopicViewer. A descrição detalhada da aplicação TopicViewer é apresentada no Anexo 4 (pág. 95).

Nesse tipo de modelagem, tópicos podem ser considerados temas sobre um determinado conjunto de documentos. Uma das abordagens para facilitar a compreensão e apresentação de resultados é a rotulação manual dos tópicos com base em suas palavras mais frequentes assim como seus documentos mais frequentes. No trabalho de Asmussen e Moller (45), em que se utilizou o algoritmo LDA, a rotulação manual dos tópicos foi realizada a partir da identificação pelos autores das dez palavras mais frequentes e também dos 20 títulos de artigos mais frequentes para cada tópico. Aqui neste estudo foi necessário realizar uma adaptação do método utilizado por esses autores em função da modelagem autor-tópico. Foi realizada a identificação das dez palavras mais frequentes e de todos os títulos de artigos dos cinco autores mais frequentes para cada tópico. Após esta identificação o próprio pesquisador realizou, a partir de uma abordagem indutiva, a rotulação dos tópicos para cada período. Para cada tópico, como resultado da rotulação, ou o tópico foi rotulado ou o tópico foi marcado como não identificado (NI). Após a rotulação dos tópicos o conjunto de tópicos de cada período foi revisado e homologado pelo orientador do próprio pesquisador.

Por meio da modelagem obtida com seus tópicos rotulados foi possível realizar uma análise exploratória que possibilitou responder a associações como: quais são os autores com maior relevância no que diz respeito a sua distribuição de frequência em determinado tópico em relação aos demais autores; se a similaridade

entre autores com base nas suas respectivas distribuições de frequência de tópicos indica uma real similaridade de seus interesses científicos.

3.8 Etapa 3: Análise de Redes de Coautorias

Para esta etapa foram realizados os seguintes passos: criação de uma rede de coautoria para cada período de cinco anos, cálculo de métricas globais e locais de rede, e análise exploratória da rede (Figura 7).

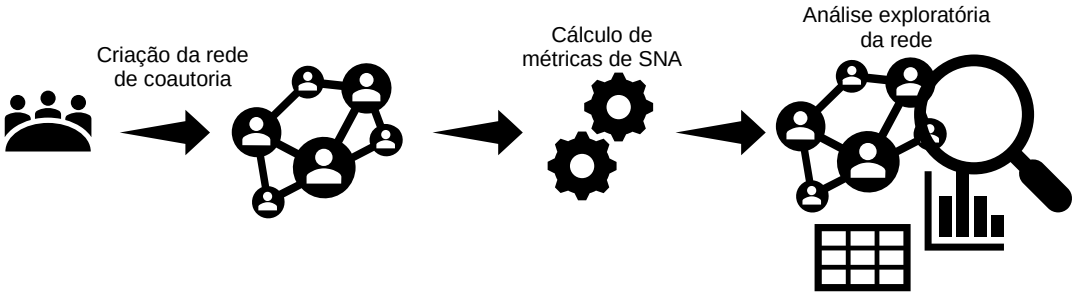


Figura 7 – Passos realizados na etapa 3 de análise de rede de coautorias.

Em análise de redes sociais, inicialmente precisamos ter conhecimento sobre a estrutura da rede que iremos analisar. Ou seja, o que serão representados como nós e o que serão representados pelas ligações. Neste trabalho foi analisada uma rede de coautoria de artigos científicos oriundos do PubMed. Assim, cada coautor foi representado por um nó e cada par de coautores que colaboraram em um mesmo artigo foi representado como uma ligação. Ainda a rede é considerada como não direcionada, o que para uma rede de coautoria indica que as ligações não possuem um determinado sentido, ou seja, se o coautor A colaborou com o coautor B, o inverso é verdadeiro. Em resumo, os dois autores colaboraram entre si.

Para a montagem das redes de coautoria, uma lista de coautorias dos artigos para cada período de cinco anos foi submetida ao software R utilizando o pacote igraph.

A partir das redes obtidas para cada período foram calculadas métricas globais de caracterização da rede:

- Número de autores: quantidade de nós ou tamanho da rede.
- Número de coautores: a quantidade de coautores(nós) que possuem ao menos um artigo com mais de um coautor além de si próprio;

- Número de coautorias: a quantidade de pares de coautores(ligações) que colaboraram em um mesmo artigo;
- Densidade da rede
- Grau médio: No caso de uma rede de coautoria, o grau médio representa com quantos coautores um coautor publicou artigos conjuntamente.
- Caminho médio
- Diâmetro da rede
- Componente gigante

Além das métricas globais para caracterização das redes, as seguintes métricas de rede locais de centralidade foram consideradas:

- Grau: a quantidade de autores que um autor possui coautoria;
- Centralidade de intermediação: a quantidade de caminhos entre pares de autores que um determinado autor faz parte.

3.9 Limitações do Estudo

Como fonte de dados somente o PubMed foi considerado. Esta escolha foi feita por esta ser a principal base de artigos na área da saúde. Outras fontes como Web of Science, Scopus, Scielo não foram consideradas neste estudo.

Para este estudo somente coautorias de artigos científicos foram consideradas como colaboração científica. Outros tipos de relações científicas como orientador-orientando, participação em um mesmo grupo de pesquisa, trabalhar em um mesmo departamento de pesquisa, entre outros tipos de colaboração possíveis não fizeram parte deste trabalho.

O conjunto de dados utilizado neste trabalho foi fechado, ou seja, coautorias entre autores aqui identificados com outros autores em artigos fora deste conjunto de dados e/ou em outras fontes/áreas do conhecimento não foram consideradas.

Não foi considerado neste estudo um tratamento de desambiguação de nomes de autores homônimos nem de deduplicação de diferentes formas da escrita ou abreviação de nomes que se referem ao mesmo autor. Hussain e Asghar (46) afirmam em seu trabalho de revisão sobre métodos de desmbiguação de nomes de autores que este é um dos problemas mais difíceis enfrentados por pesquisadores de bases de dados digitais. Em seu artigo de revisão da literatura sobre redes de coau-

toria, Kumar (6) salienta que a desambiguação de nomes ainda é um problema não resolvido. Alguns trabalhos evitam sua utilização enquanto outros indicam um método, mas não detalham a solução aplicada. Diversas abordagens de tratamento da desambiguação são encontradas na literatura. Como exemplo, Kang e colaboradores (47) utilizaram uma abordagem que adota como premissa que a identidade de um autor pode ser determinada por seus coautores. Já Strotmman e colaboradores (48) utilizaram uma abordagem baseada na similaridade entre diferentes publicações que constam o mesmo nome de autor.

4 RESULTADOS

4.1 Etapa 1: Coleta de Artigos do PubMed

No primeiro passo foram identificados 69 periódicos que atenderam a estratégia de busca definida na base NLM Catalog. O Quadro 2 apresenta os nomes abreviados de todos os periódicos identificados.

Quadro 2 – Lista de periódicos obtida do NLM Catalog.

Periódicos	
AMIA Annu Symp Proc	Int J Neural Syst
Appl Clin Inform	J AHIMA
Artif Intell Med	J Am Med Inform Assoc
Big Data	J Biomed Inform
BMC Med Inform Decis Mak	J Biomed Semantics
Brief. Bioinformatics	J Chem Inf Model
Comput Biol Chem	J Clin Monit Comput
Comput Inform Nurs	J Digit Imaging
Comput Math Methods Med	J Healthc Eng
Comput Methods Programs Biomed	J Innov Health Inform
Comput Syst Bioinformatics Conf	J Integr Bioinform
Comput. Biol. Med.	J Med Syst
Curr Comput Aided Drug Des	J Telemed Telecare
Curr Protoc Bioinformatics	J. Comput. Biol.
Geospat Health	J. Med. Internet Res.
Gigascience	Lifetime Data Anal
Health Info Libr J	Med Biol Eng Comput
Health Informatics J	Med Decis Making
Health Manag Technol	Med Image Comput Comput Assist Interv
Healthc (Amst)	Med Ref Serv Q
Healthc Inform	Medinfo
HIM J	Methods Inf Med
IEEE Comput Graph Appl	Neural Comput
IEEE J Biomed Health Inform	Neuroinformatics
IEEE Trans Image Process	Perspect Health Inf Manag
IEEE Trans Neural Netw Learn Syst	Prog Community Health Partnersh

IEEE Trans Pattern Anal Mach Intell	Res Synth Methods
IEEE Trans Vis Comput Graph	Sci Data
Inf Process Med Imaging	Spat Spatiotemporal Epidemiol
Inform Health Soc Care	Stud Health Technol Inform
Int J Comput Assist Radiol Surg	Teach Learn Med
Int J Comput Dent	Telemed J E Health
Int J Health Geogr	Wiley Interdiscip Rev Syst Biol Med
Int J Med Inform	Yearb Med Inform
Int J Med Robot	

No segundo passo foram coletados do PubMed os artigos dos 69 periódicos identificados no passo anterior, a quantidade de periódicos com artigos publicados por ano de publicação pode ser observada na Figura 8.

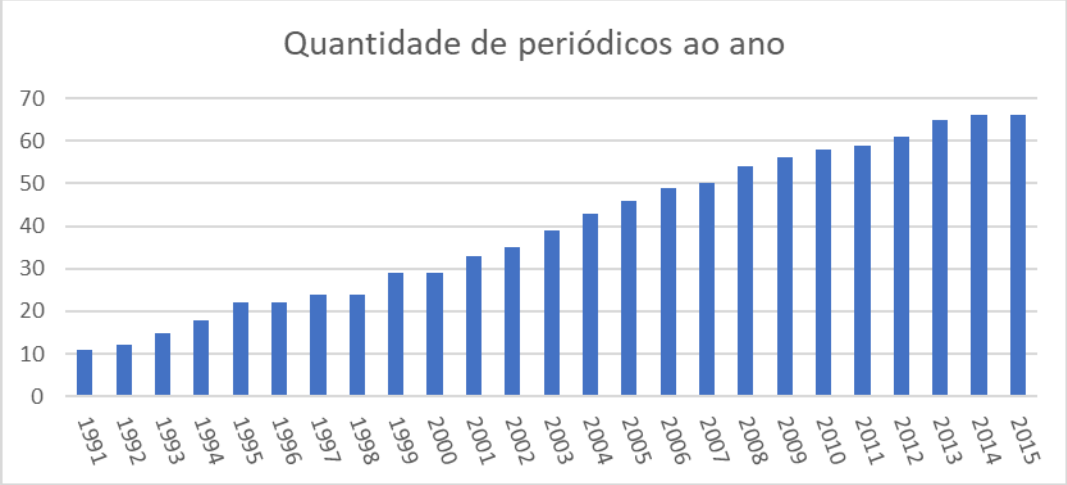


Figura 8 – Quantidade de periódicos identificados por ano de publicação.

O total de artigos coletados do PubMed publicados nestes periódicos foi de 76.250. Na Figura 4 é apresentada a quantidade de artigos coletados por ano de publicação. A lista completa de artigos pode ser acessada em (https://bit.ly/39sw9Me). Os artigos foram agrupados em períodos de cinco anos conforme Tabela 1.

Período	Artigos
1991-1995	3.475
1996-2000	7.183
2001-2005	12.123
2006-2010	21.849
2011-2015	31.620

Tabela 1 – Número de artigos obtidos por períodos de cinco anos.

Figura 9 – Quantidade de artigos por ano de publicação.

4.2 Etapa 2: Modelagem Autor-Tópico

Após a execução do pré-processamento, um conjunto de dados para cada período de cinco anos foi obtido. Na Tabela 2 são apresentados os totais obtidos por período.

	1991-1995	1996-2000	2001-2005	2006-2010	2011-2015
Número de periódicos em IS	22	30	43	60	68
Número de artigos	3.475	7.183	12.123	21.849	31.620
Número de palavras únicas	1.285	1.345	1.392	1.452	1.627
Número de autores únicos	7.561	15.782	31.257	55.883	92.355
Média de autores por artigo	2,92 +-1,90	3,20+-2,04	3,58+-2,29	3,90+-2,39	4,44+-3,02

Tabela 2 - Resultados do pré-processamento.

Na etapa seguinte foi obtido um modelo AT com 50 tópicos para cada período de cinco anos.

O tempo de processamento dos modelos podem ser observados na Tabela 3. A complexidade de tempo de cada iteração do amostrador de Gibbs para o modelo AT é relacionada diretamente ao número total de palavras do corpus multiplicado pelo número de tópicos (k). O número de coautores em cada artigo tem pouca influência no tempo de processamento conforme observado por Rosen-Zvi e colabo-

radores (17). O processamento dos modelos foi realizado em um notebook com processador Intel® I7® 5500U (2.40 GHz, 64-bit), 8 GB de memória RAM.

Período	Tempo de Processamento (hh:mm)
1991-1995	00:21
1996-2000	00:44
2001-2005	01:29
2006-2010	02:23
2011-2015	04:12

Tabela 3 – Tempos de processamento dos modelos para cada período.

As representações dos modelos AT podem ser acessados dinamicamente por meio dos endereços eletrônicos:

- 1991-1995: http://atviewer1.azurewebsites.net/

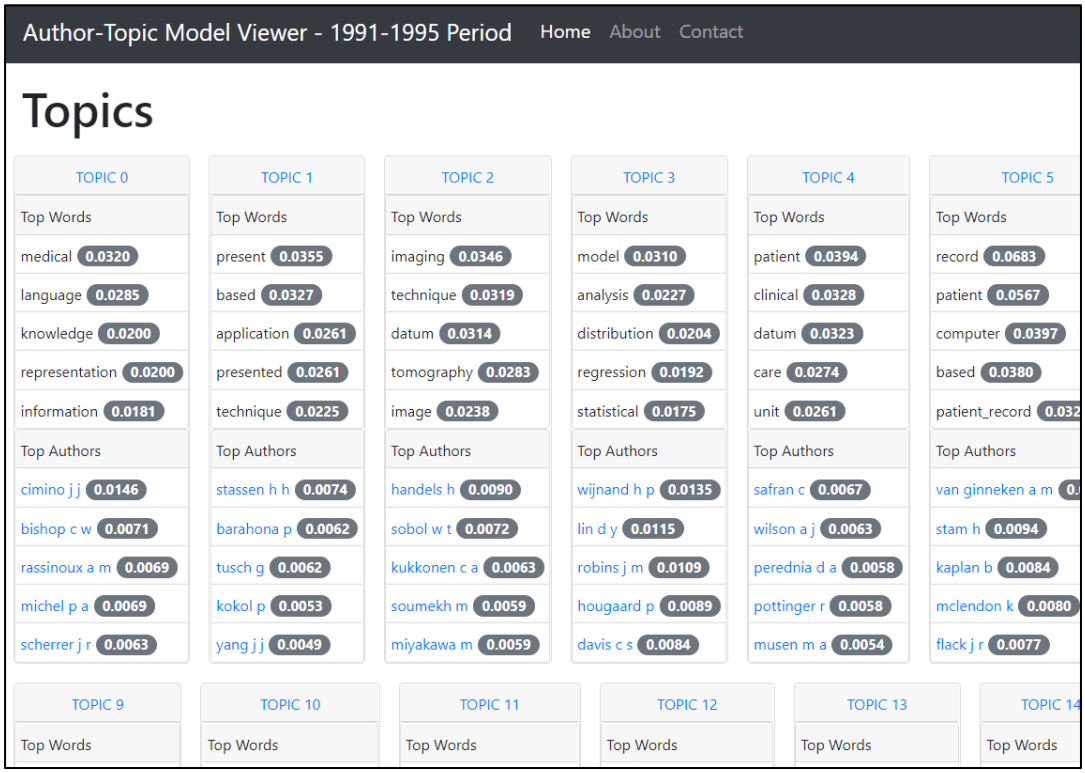


Figura 10 – Recorte da tela de abertura do visualizador do modelo 1991-1995. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico.

- 1996-2000: http://atviewer2.azurewebsites.net/

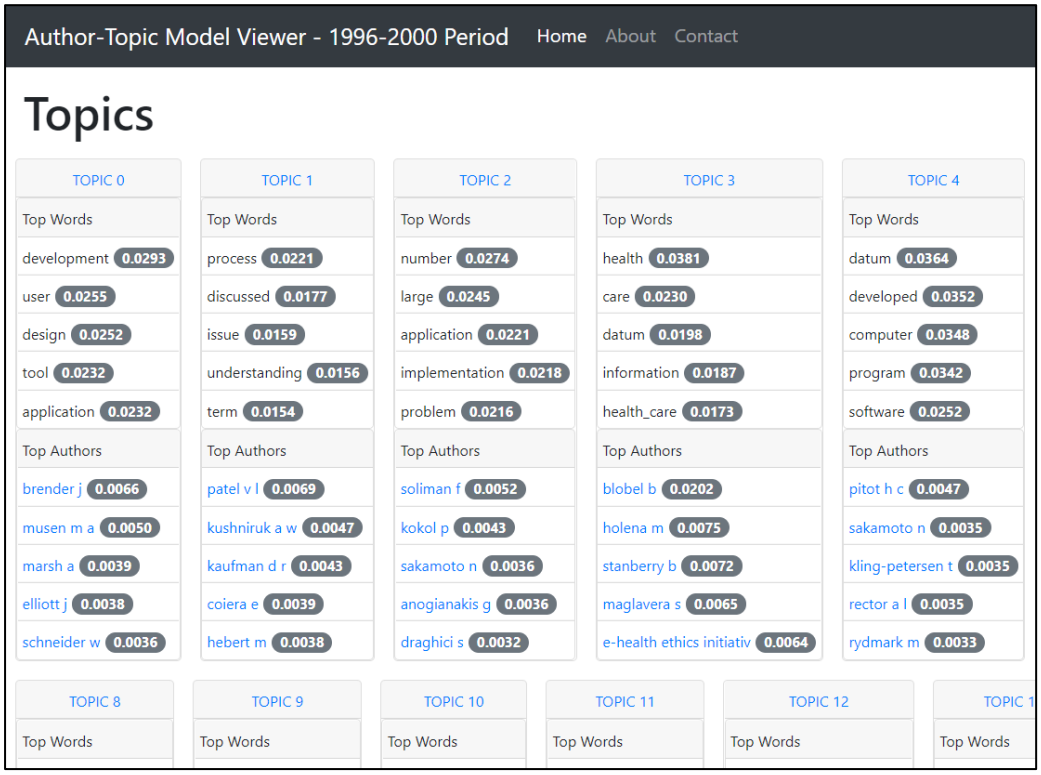


Figura 11 - Recorte da tela de abertura do visualizador do modelo 1996-2000. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico.

- 2001-2005: http://atviewer3.azurewebsites.net/

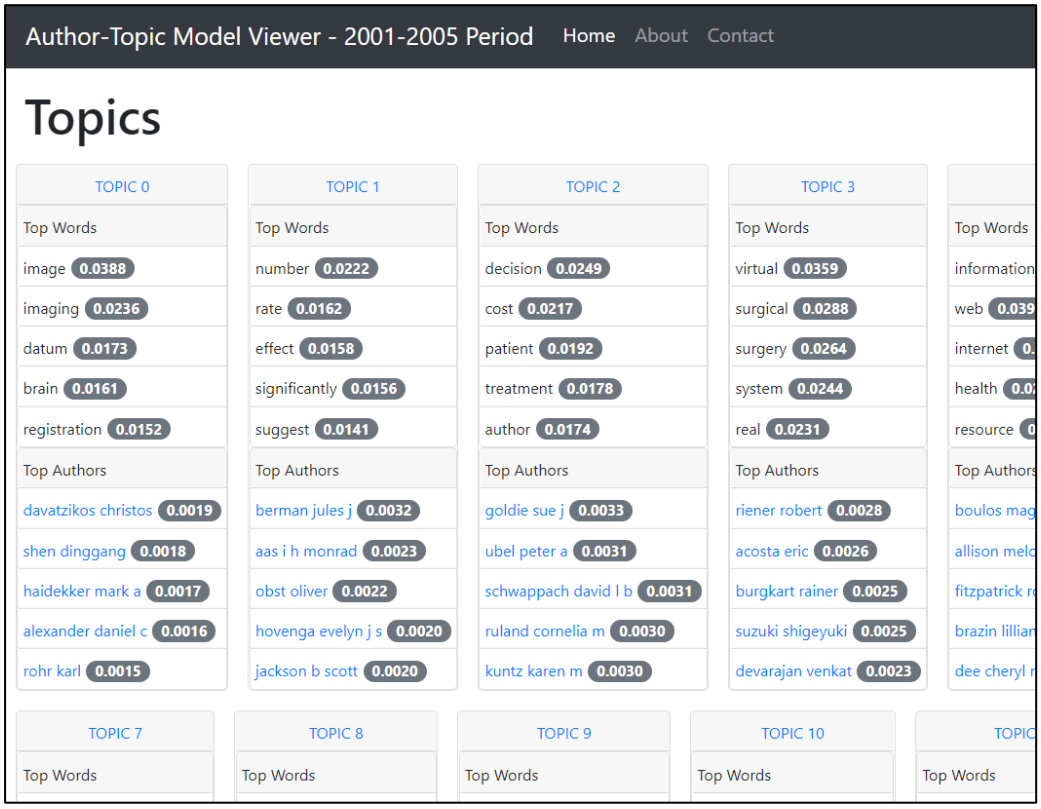


Figura 12 - Recorte da tela de abertura do visualizador do modelo 2001-2005. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico.

- 2006-2010: http://atviewer4.azurewebsites.net/

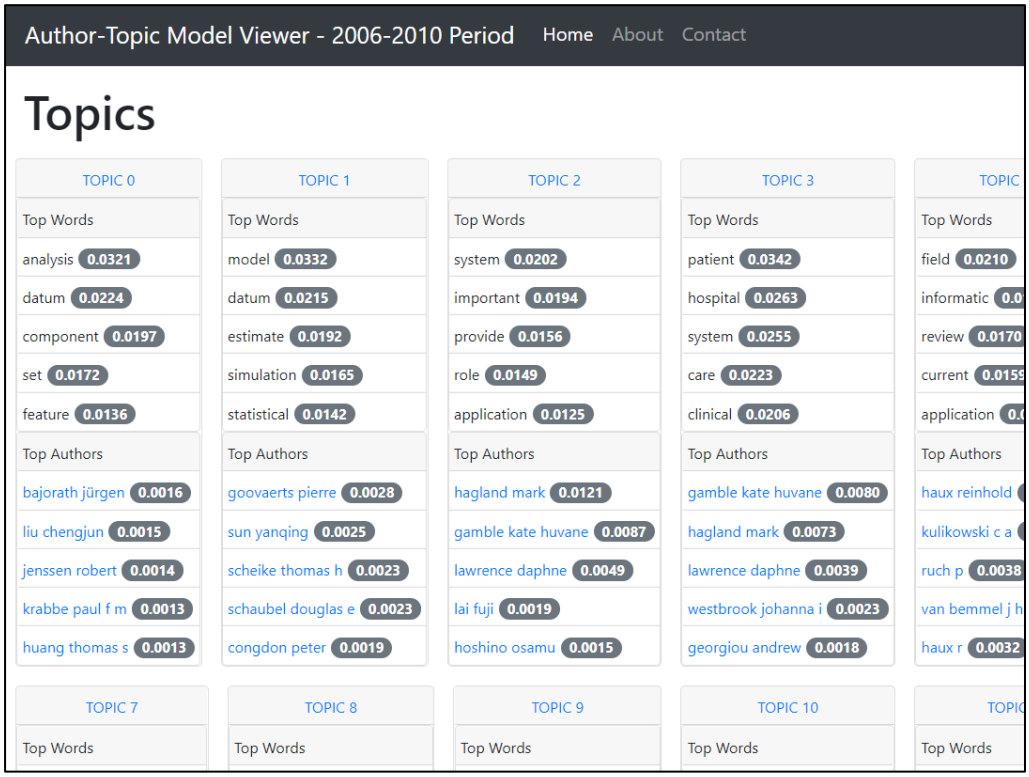


Figura 13 - Recorte da tela de abertura do visualizador do modelo 2006-2010. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico.

- 2011-2015: http://atviewer5.azurewebsites.net/

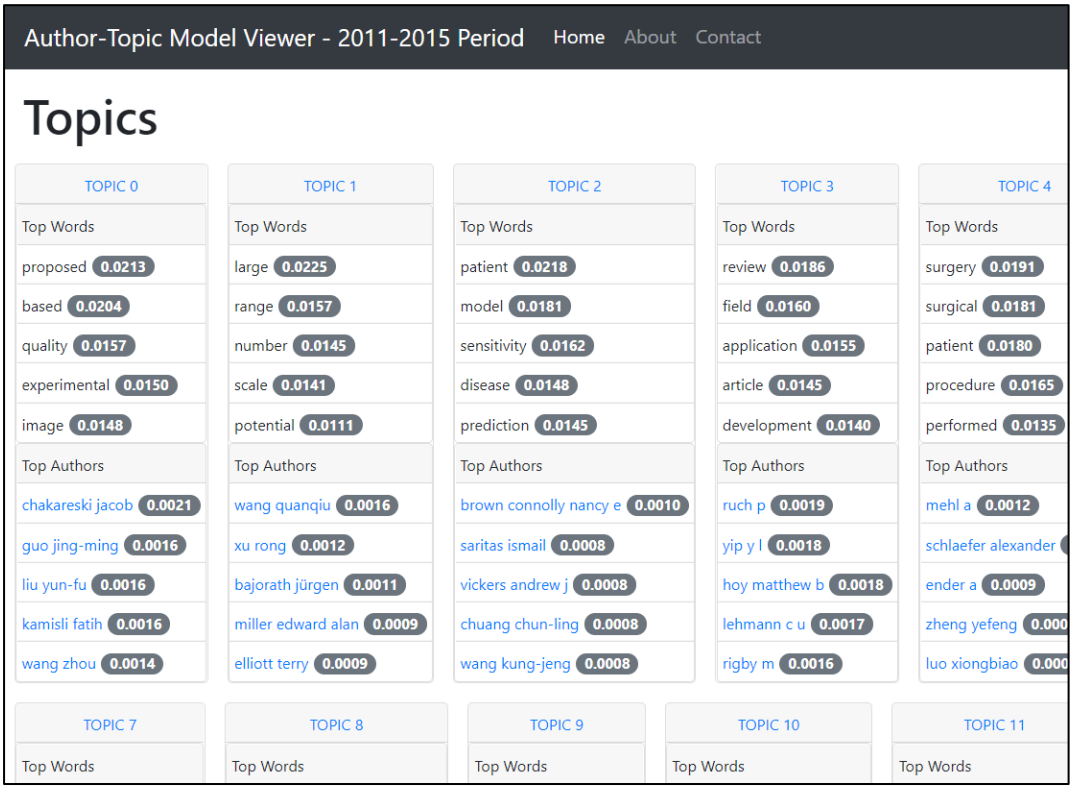


Figura 14 - Recorte da tela de abertura do visualizador do modelo 2011-2015. A tela de abertura apresenta os 50 tópicos com as cinco palavras mais frequentes e os cinco autores mais frequentes para cada tópico.

Para cada modelo, cada tópico foi representado pelas cinco palavras com maior frequência e os cinco autores com maior frequência para o tópico. Como exemplo, na Figura 15 são apresentados os tópicos 0, 7 e 31 do modelo do período de 1991 a 1995. Ao lado de cada palavra é apresentada a frequência com que esta ocorre no tópico, assim como ao lado de cada autor é apresentada a frequência com que este ocorre no tópico. Ambas frequências são relativas, variando entre 0 e 1.

Figura 15 – Tópicos 0, 7 e 31 do período de 1991 a 1995

Cabe salientar ainda que cada tópico de um modelo possui a mesma lista de palavras e a mesma lista de autores, o que os diferencia são as distribuições de frequência das palavras e dos autores. Ou seja, do ponto de vista prático, a ordem das palavras e dos autores é alterada em cada tópico. Assim, ao se aplicar um corte considerando um valor mínimo de frequência pode-se considerar como sendo grupos diferentes de palavras e autores em cada tópico. (tem que ver se desse jeito fica bem claro)

4.2.1 Rotulação dos tópicos obtidos

A rotulação manual de tópicos foi apoiada pelo uso da aplicação TopicViewer para cada um dos cinco períodos analisados tanto para a rotulação quanto para a revisão e homologação pelo orientador da pesquisa. Os tópicos rotulados e os que não foram possíveis de serem rotulados se encontram no anexo 5 (pág. 100).

Na Tabela 4 são apresentados o número de tópicos que foram possíveis de ser interpretados e rotulados por período analisado.

Período	Número de Tópicos
1991-1995	33(66%)
1996-2000	36(72%)
2001-2005	37(74%)
2006-2010	43(86%)
2011-2015	42(84%)

Tabela 4 – Número de tópicos identificados por período.

4.2.2 Análise exploratória de cada período de cinco anos

- Período entre os anos de 1991 e 1995

Neste período foram identificados 33 dos 50 tópicos obtidos. Dos 33 tópicos possíveis de serem rotulados, foi observado que vários tópicos discorrem sobre um mesmo grande tema, porém cada tópico com sua especificidade. Por exemplo, seis tópicos foram rotulados como subtemas de processamento de imagens médicas, cada tópico com seu subtema específico. Por exemplo o tópico 10 discorre sobre compressão de imagens médicas, já o tópico 23, filtros em imagens médicas e o tópico 25, sobre qualidade de imagens (Figura 16).

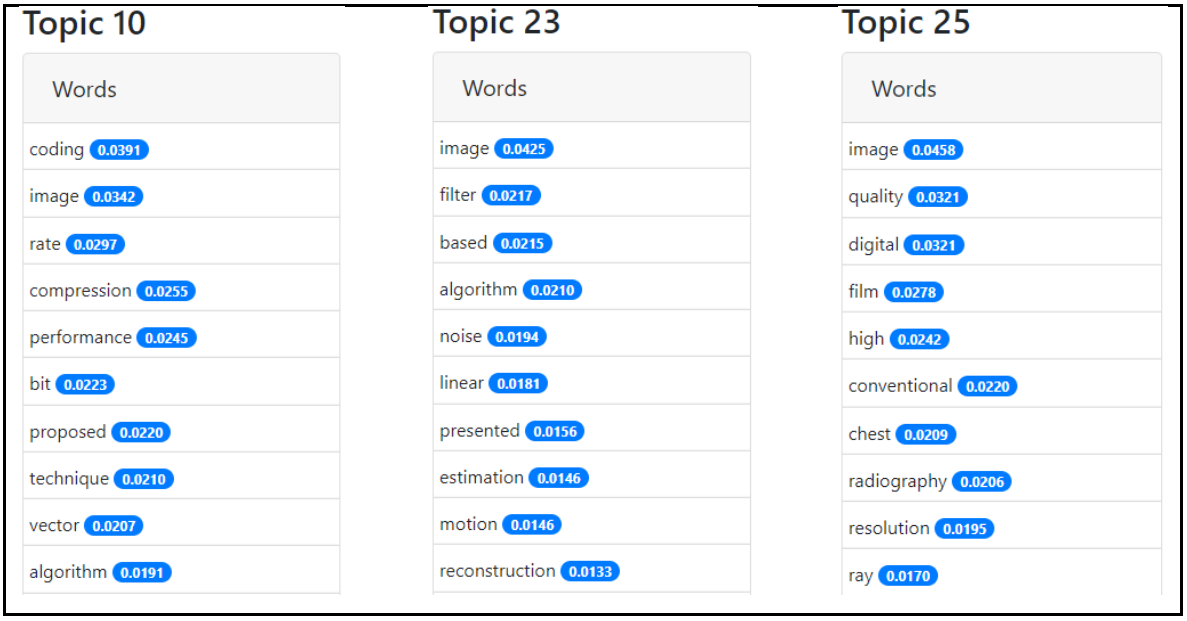


Figura 16 – Tópicos sobre processamento de imagens e sinais médicos.

Também pode ser observada a presença de importantes autores para literatura de IS como Edward H. Shortliffe (shortliffe e h), Charles Safran (safran c) e James J Cimino (cimino j j), que por exemplo é o autor com maior probabilidade para o tópico 0, rotulado como ‘representação do conhecimento médico’. Este resultado corrobora com o interesse do autor no período com vários artigos sobre o tema no período. Na Figura 17 são apresentadas os artigos do autor coletados para o período.

cimino j j

Publications
IAIMS and UMLS at Columbia-Presbyterian Medical Center. - Med Decis Making - 1991
Using the UMLS to bring the library to the bedside. - Med Decis Making - 1991
Automatic knowledge acquisition from MEDLINE. - Methods Inf Med - 1993
Knowledge-based approaches to the maintenance of a large controlled medical terminology. - J Am Med Inform Assoc - 1994
A general natural-language text processor for clinical radiology. - J Am Med Inform Assoc - 1994
Toward a medical-concept representation language. The Canon Group. - J Am Med Inform Assoc - 1994
A schema for representing medical language applied to clinical radiology. - J Am Med Inform Assoc - 1994
Coding Systems in Health Care. - Yearb Med Inform - 1995
Medical Informatics Training at Columbia University and the Columbia-Presbyterian Medical Center. - Yearb Med Inform - 1995
Internet as clinical information system: application development using the World Wide Web. - J Am Med Inform Assoc - 1995
The Canon Group's effort: working toward a merged model. - J Am Med Inform Assoc - 1995
Managing vocabulary for a centralized clinical system. - Medinfo - 1995
Use of the Unified Medical Language System in patient care at the Columbia-Presbyterian Medical Center. - Methods Inf Med - 1995

Figura 17 – Lista de artigos do autor James J. Cimino.

Outro achado importante foi o tópico 41, que não foi possível rotular. As palavras mais frequentes deste tópico não indicam claramente um assunto. Ao navegar pelos autores e seus respectivos artigos e resumos, podemos observar que não há uma clara ligação entre estes e o tópico. Este achado pode indicar que as palavras mais relevantes deste tópico são candidatas a serem removidas quando do processo inicial de limpeza de corpus, com objetivo de melhorar o resultado em um novo processamento do modelo.

- Período entre os anos de 1996 e 2000

Neste período foram rotulados 36 tópicos dos 50 tópicos obtidos. Assim como no período anterior foram observados vários tópicos do mesmo assunto. Como exemplo, o tópico “processamento de imagens e sinais médicos” pode ser associado com sete tópicos, tendo cada um a sua especificidade.

Novamente foi possível observar autores consagrados no período como Vimla Patel (patel v), John Mantas (mantas j), Lucila Ohno-Machado (ohno-machado l) e Ricardo Bellazzi (bellazzi r).

Uma observação interessante para este período foi o tópico 15, denominado “informações de saúde na web”, corroborando com o período de disseminação do uso da rede mundial de computadores (*world wide web, www*) criada por Tim Berners-Lee em 1989 (49).

Figura 18 – Tópico 15 - “informações de saúde na web”.

Embora com origens bem anteriores a este período, o tópico 33 (Figura 19) foi claramente identificado como “telemedicina”, destacando-se no período, o periódico *Journal of Telemedicine and Telecare* lançado em 1995, periódico com maior fator de impacto da área(50).

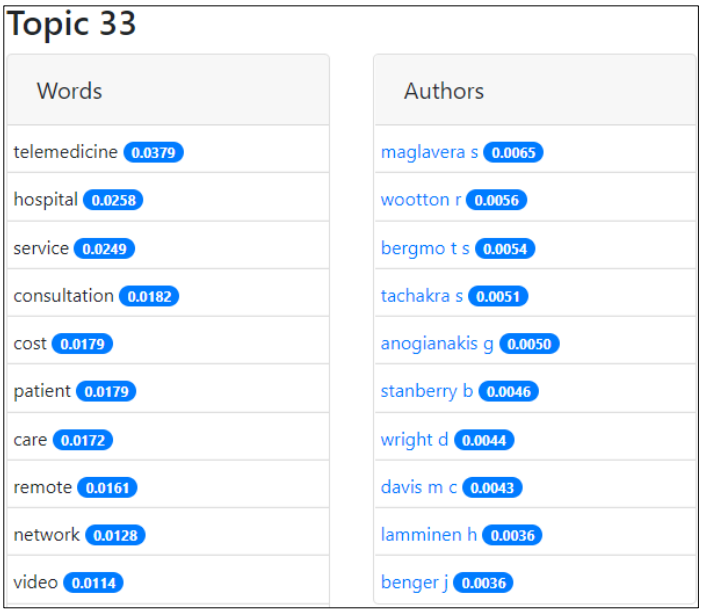


Figura 19 – Tópico 33, rotulado como Telemedicina, as palavras e os autores com maior probabilidade para o tópico.

Este resultado também corrobora com o forte interesse em telemedicina pelo autor Richard Wooton (wooton r), uma referência na área de telemedicina com 24 artigos no período (Figura 20).

Figura 20 – Recorte de 10 dos 24 artigos no período de 1996-2000 do autor Richard Wooton.

Outro tópico que emerge no período é o tópico 46 (Figura 21) que discorre sobre “simulação e treinamento por meio de realidade virtual”.

Figura 21 – Tópico 46 sobre simulação e treinamento por meio de realidade virtual.

- Período entre os anos de 2001 e 2005

Neste período emerge o tópico 18 (Figura 22), rotulado como “saúde pública”, com destaque para o periódico International Journal of Health Geographics cujo escopo relaciona entre outros, temas relacionados a saúde pública com apoio de sistemas de informação geográficas.

Figura 22 – Tópico 18 sobre saúde pública.

Outro tópico emergente no período foi o tópico 45 (Figura 23), rotulado como “quimioinformática”. Este achado foi fortemente influenciado pelo aumento significa-

tivo do número de pesquisas, o que gerou um reconhecimento da área neste período (51).

Figura 23 – Tópico 45 sobre quimioinformática.

- Período entre os anos de 2006 e 2010

Neste período há um destaque para o tópico 23 (Figura 24) rotulado como “cirurgia assistida por robô”. Além das palavras mais prováveis identificarem com clareza a área de pesquisa, os autores mais prováveis são referências no tema como por exemplo o Prof. Dr. Guoyan Zheng (zhen guoyan), diretor do Institute of Medical Robotics da ShanghaiJiao Tong University (https://orcid.org/0000-0003-4173-0379).

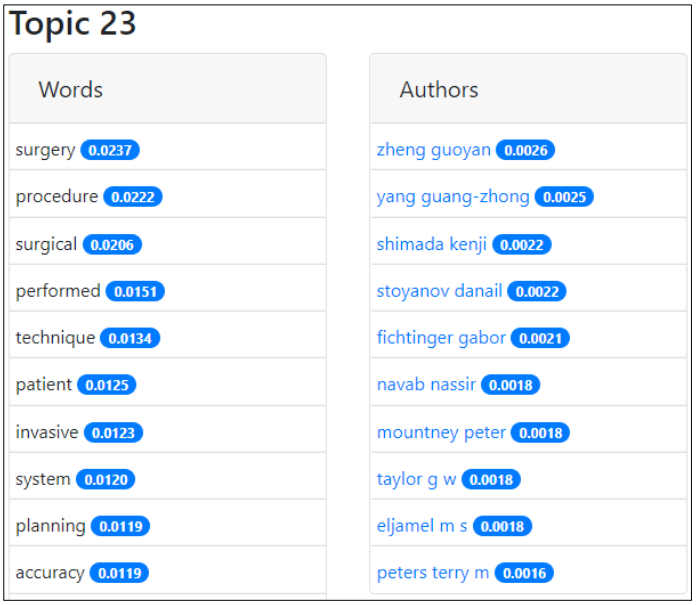


Figura 24 – Tópico 23 sobre cirurgia assistida por robô

Também cabe destacar o tópico 41 (Figura 25) rotulado como “usabilidade de software”.

Figura 25 – Tópico 41 – “usabilidade de software”

- Período entre os anos de 2011 e 2015

Neste período, cabe destacar o tópico 33 (Figura 26) rotulado como “segurança da informação em saúde”. Cabe salientar que, os autores mais prováveis para este tópico, também tem este tópico com a frequência mais alta entre os 50 tópicos. Sendo esta frequência acima de 70% para cada um dos autores, demonstrando uma forte especificidade do tema pelos autores.

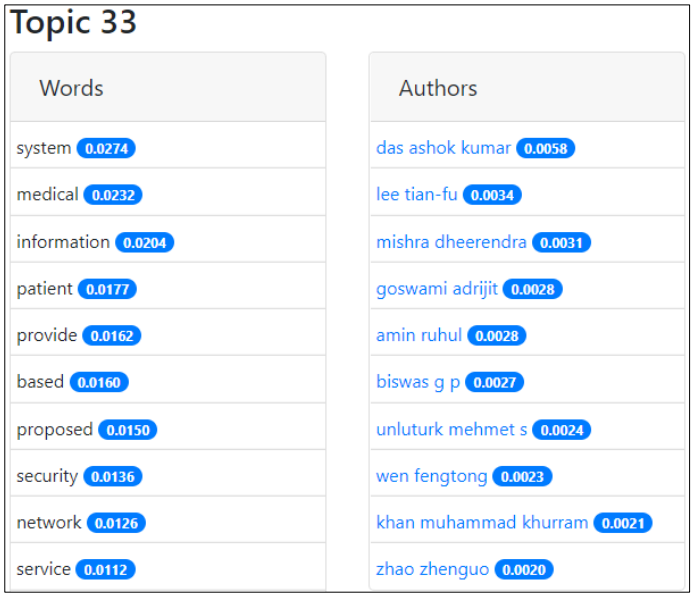


Figura 26 – Tópico 33 – “segurança da informação em saúde”

4.3 Etapa 3: Análise de Redes de Coautorias

Como primeiro passo desta etapa foram obtidas cinco redes de coautoria não direcionadas, uma para cada período de 5 anos: 1991-1995, 1996-2000, 2001-2005, 2006-2010 e 2011-2015. As métricas obtidas para cada rede são apresentadas na Tabela 5.

Métrica	1991-1995	1996-2000	2001-2005	2006-2010	2011-2015
Núm. de autores	7.561	15.782	31.257	55.883	92.355
Núm. de autores sem coautoria	538(7,11%)	720(4,56%)	997(3,19%)	963(1,72%)	812(0,88%)
Número de coautores	7.023	15.062	30.260	54.920	91.543
Número de coautorias	14.242	35.094	78.253	162.212	342.907
Grau médio	3,76	4,45	5,00	5,81	7,43
Densidade	0,0006	0,0003	0,0002	0,0001	0,00008
Diâmetro	20	28	28	30	28
Caminho médio	8,24	11,05	7,73	9,30	8,33
Tamanho componente Gigante	1.321	3.871	3.939	23.089	47.482
% componente gigante	18,81%	25,70%	13,02%	42,04%	51,87%

Tabela 5 – Métricas globais de caracterização das redes.

Para o caminho médio também foi calculada a distribuição de tamanhos de caminhos, o resultado pode ser observado na Figura 27.

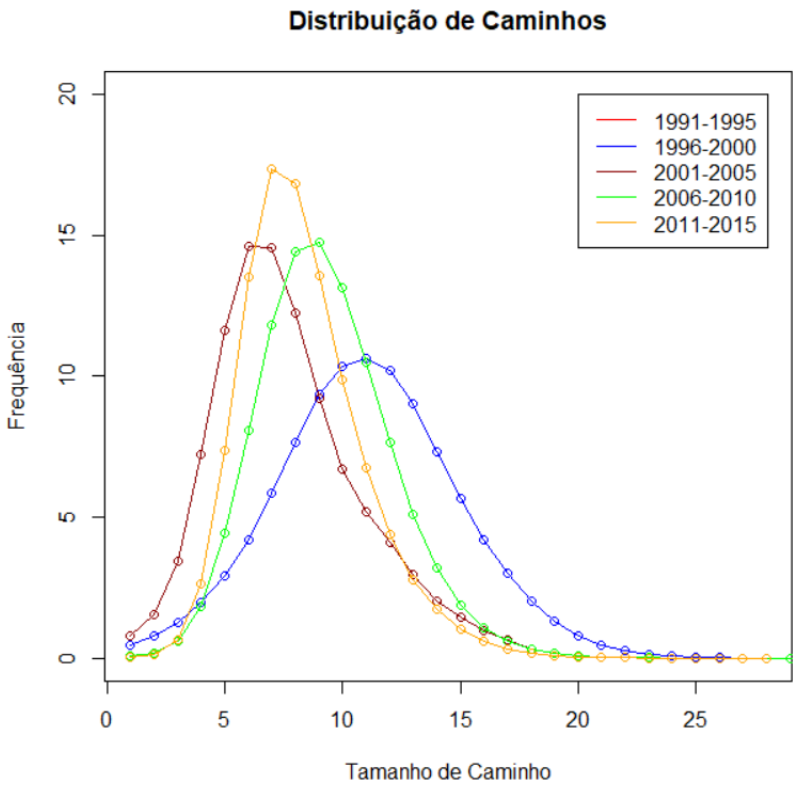


Figura 27 – Distribuição de frequências (%) dos tamanhos de caminho.

Para o grau médio também foi calculada a distribuição de grau, como pode ser observado pela Figura 28.

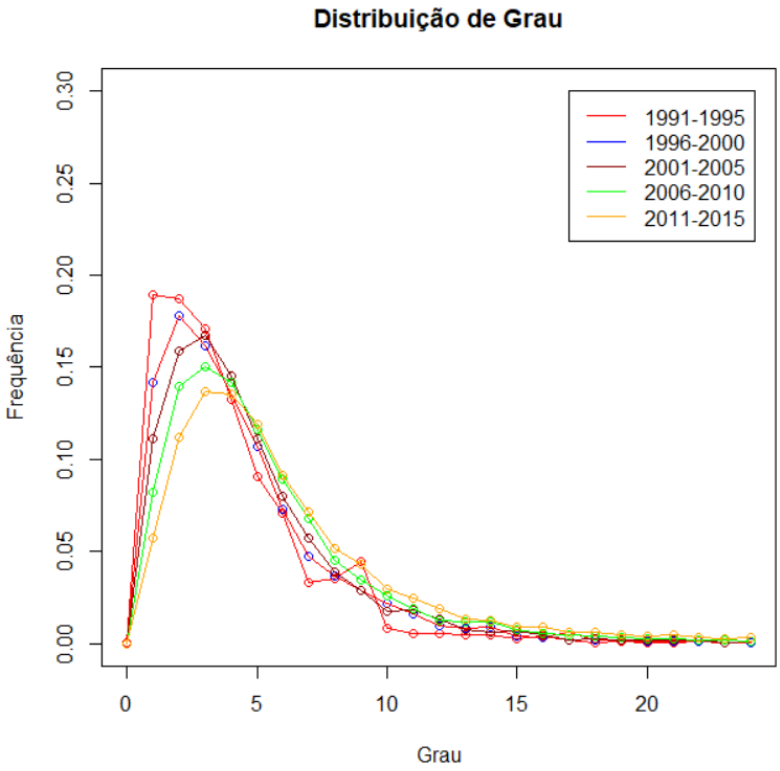
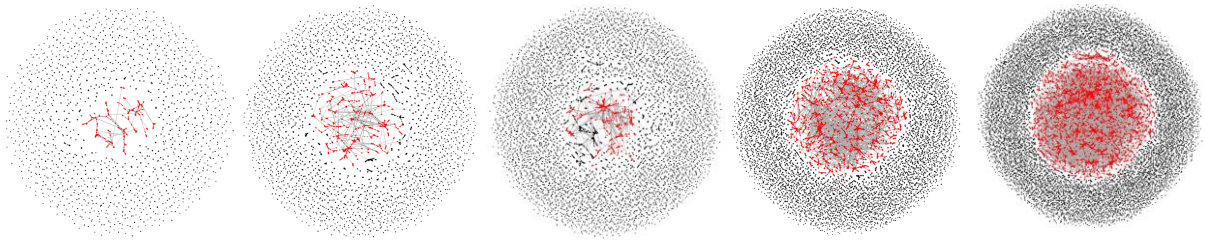


Figura 28 – Distribuição de frequencias (%) de grau.

Na Figura 29 é apresentada uma visualização das redes de cada período, com destaque ao centro e em vermelho para os autores que fazem parte do componente gigante, ou seja, a subrede que possui pelo menos um caminho entre todos os pares de autores da rede.



(a) (b) (c) (d) (e)

Figura 29 - Representação gráfica das redes obtidas em cada período: (a) 1991 à 1995; (b) 1996 à 2000; (c) 2001 à 2005; (d) 2006 à 2010; (e) 2011 à 2015.

Em seguida foram calculadas as métricas locais de grau, grau poderado e centralidade de intermediação para caracterização individual de cada autor. A Tabela 6 apresenta os cinco autores com maior grau em cada período.

Período	Autor	Grau
1991-1995	Wigertz O	45
	Degoulet P	43
	Takahashi T	40
	Scherrer J R	32
	Fagan L M	31
1996-2000	Haux R	74
	Wootton R	67
	Takahashi T	50
	Hasman A	45
	Cerutti S	45
2001-2005	Bates David W	135
	Haux R	97
	Bakken Suzanne	84
	Miller Perry L	80
	Martin-Sanchez F	77
2006-2010	Bates David W	176
	Middleton Blackford	146
	Haux Reinhold	127
	Bakken Suzanne	123
	Cimino James J	116
2011-2015	Wang Jun	342
	Bates David W	237
	Chute Christopher G	237
	Davies Neil	215
	Bicak Mesude	215

Tabela 6 – Cinco autores com maior grau em cada período.

A centralidade de intermediação (*betweeness*) foi calculada e os cinco autores com maior centralidade de intermediação são apresentados na Tabela 7.

Período	Autor	Centralidade de intermediação
1991-1995	Beck J R	1,23%
	Inoue Y	1,13%
	Saranummi N	1,07%
	Takahashi T	1,04%

	Satoh K	0,97%
1996-2000	Haux R	1,32%
	Lun K C	1,11%
	Lee S	1,09%
	Yang S	1,07%
	Wang C	1,05%
2001-2005	Overhage J Marc	0,21%
	Shortliffe Edward H	0,19%
	Miller Perry L	0,17%
	Bates David W	0,17%
	Hripcsak George	0,17%
2006-2010	Pennec Xavier	0,74%
	Hripcsak George	0,73%
	Bates David W	0,69%
	Shen Dinggang	0,56%
	Overhage J Marc	0,55%
2011-2015	Bates David W	0,71%
	Hripcsak George	0,54%
	Chute Christopher G	0,52%
	Wang Jun	0,45%
	Zheng Kai	0,43%

Tabela 7 – Cinco autores com maior centralidade de intermediação em cada período.

Após a criação das redes e cáculos das métricas globais e locais foi realizada uma análise exploratória em cada período estudado. Foram analisados os componentes gigantes de cada período, assim como os autores mais importantes em relação a centralidade. O componente gigante de uma rede de coautoria concentra os autores mais importantes da rede, embora existam exceções (9).

Para os autores com maior grau e centralidade de intermediação em cada período foram extraídas subredes conhecidas como redes egocêntricas. Em análise de redes sociais uma rede egocêntrica se caracteriza por se mapear uma rede a partir de um determinado nó denominado ego e identificando suas ligações primárias, conhecidas como alters. E a partir daí, as ligações de seus alters, e assim por diante (52). Cada nível de ligação é chamado de ordem. Ordem igual a 1 representa a rede composta do ego e seus alters apenas. Ordem igual 2 representa uma rede composta pelo ego, seus alters e os alters dos alters e assim sucessivamente para ordem 3, 4, etc. Cabe salientar que existe a ordem 1,5, em que além das ligações do ego e

seus alters, são consideradas também as ligações entre seus alters. Cabe salientar que o estudo de redes egocêntricas é uma subárea da análise de redes sociais bem ampla, e este trabalho se limitou apenas a ilustrar a composição das ligações de co-autoria dos autores em destaque.

4.3.1 Redes de Coautorias 1991-1995

No período de 1991 a 1995, o componente gigante concentrou todos os cinco autores com maior grau, grau ponderado e intermediação.

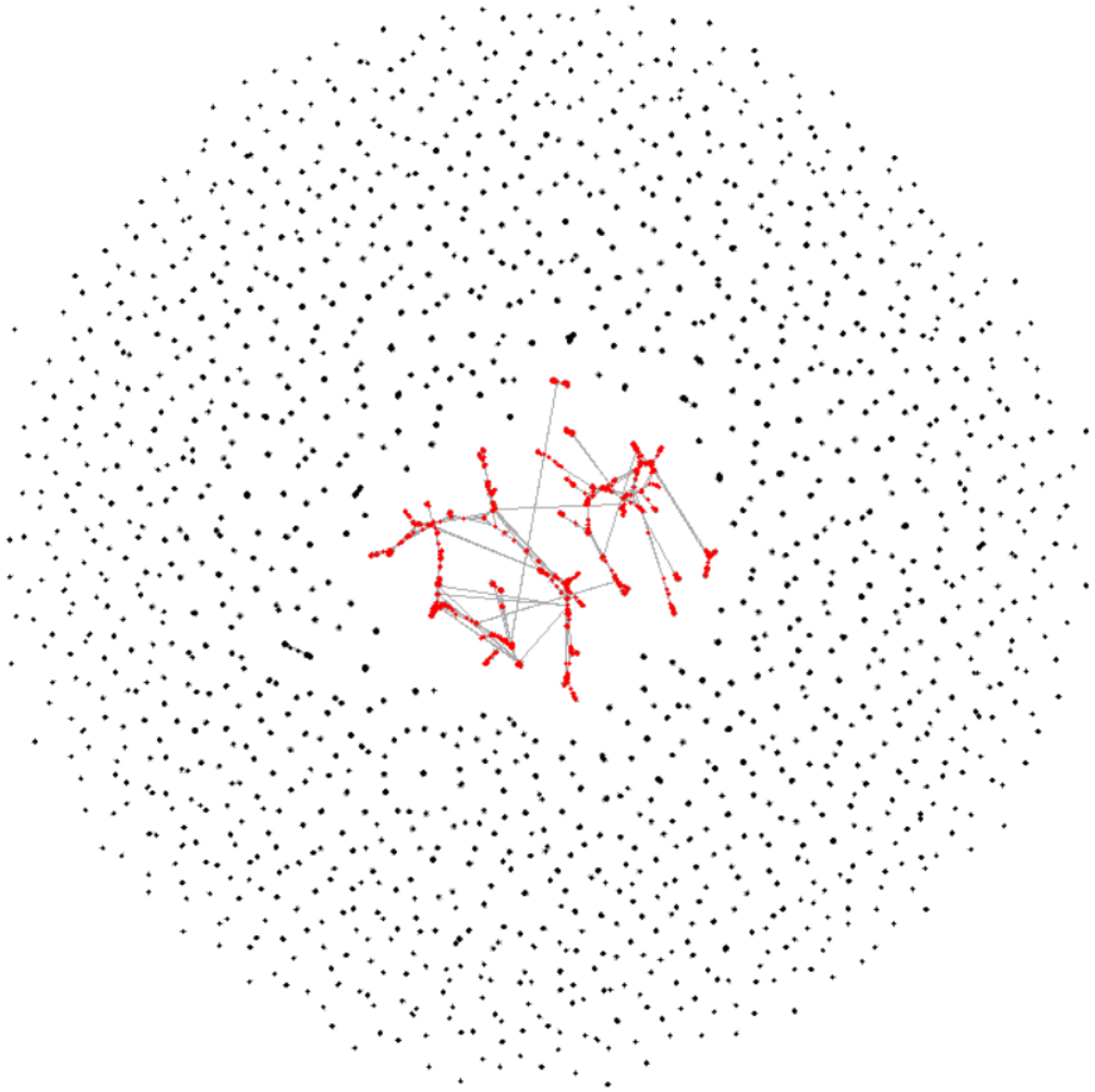


Figura 30 – Representação gráfica da rede de 1991-1995 com 7.561 autores e 14.242 coautorias. A figura em alta resolução está disponível em https://bit.ly/2HSveZU .

O autor com maior grau neste período foi o pesquisador Ove B. Wigertz, identificado neste trabalho como “Wigertz O”, com grau 45, ou seja, possui coautorias com outros 45 autores. Apenas para ilustrar sua importância, o pesquisador sueco Ove B. Wigertz foi eleito mebro “fellow” da AMIA (American Medical Informatics Association) em 1997 (https://www.amia.org/about-amia/leadership/acmi-fellow/ove-b-wigertz-dsc-dmedsc-facmi). O programa de fellowship da AMIA reconhece nomes tanto dos Estados Unidos quanto de outros países. São eleitos os pesquisadores que realizaram contribuições significativas e sustentadas para a área de informática em saúde (53). Na Figura 31 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores. Outra observação é sua coautoria com o pesquisador Patrice Degoulet que possui o segundo maior grau no período.

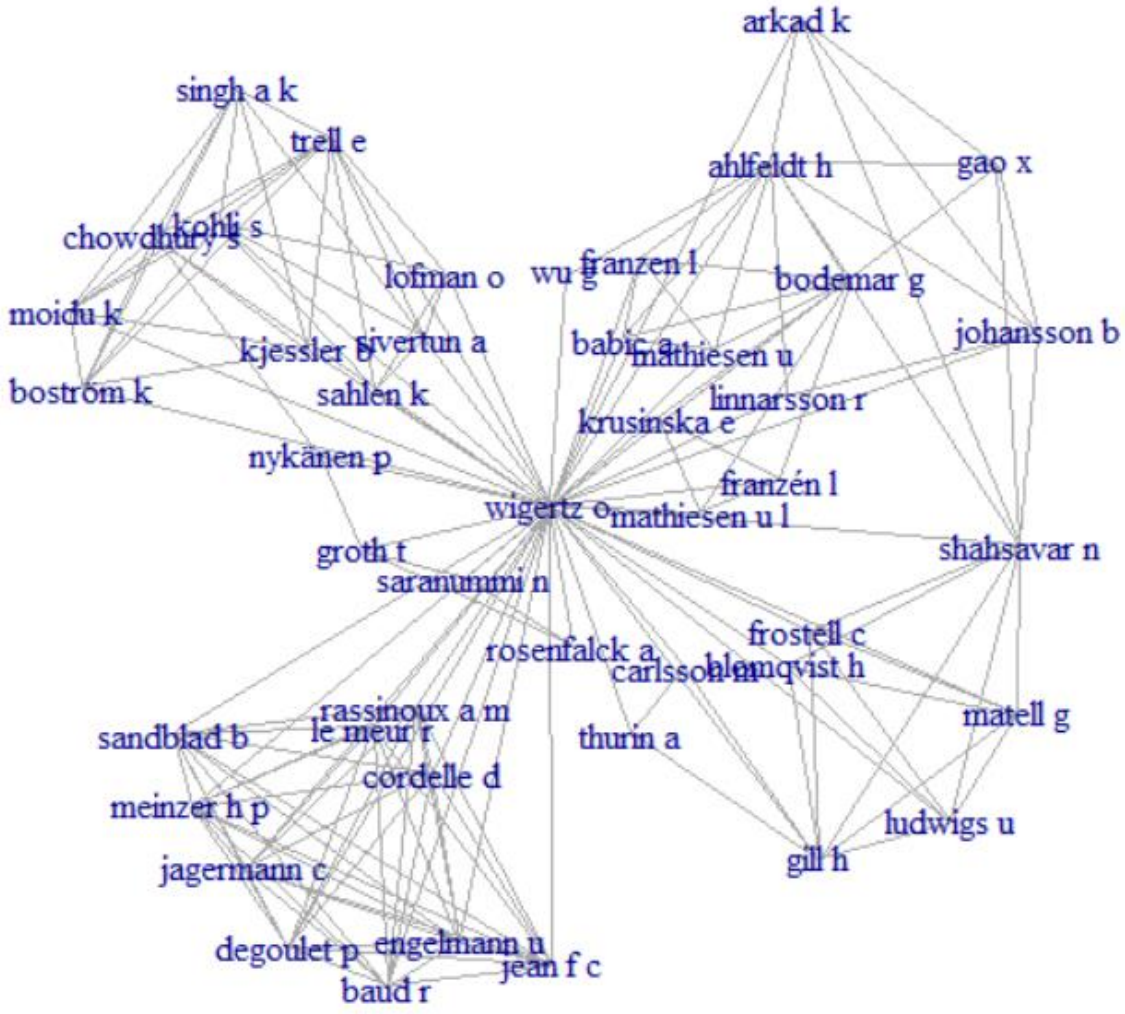


Figura 31 – Rede egocêntrica do pesquisador Ove B. Wigertz no período de 1991 a 1995.

Já em relação a centralidade de intermediação, o pesquisador J. Robert Beck, neste trabalho identificado como “Beck J R”, é o autor mais central em relação a intermediação com 1,23% dos caminhos entre outros autores da rede passando por ele. Também para ilustrar sua importância, o pesquisador J. Robert Beck foi eleito membro “fellow” da AMIA (American Medical Informatics Association) em 1990 (https://www.amia.org/about-amia/leadership/acmi-fellow/j-robert-beck-md-facmi). Na Figura 32 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores.

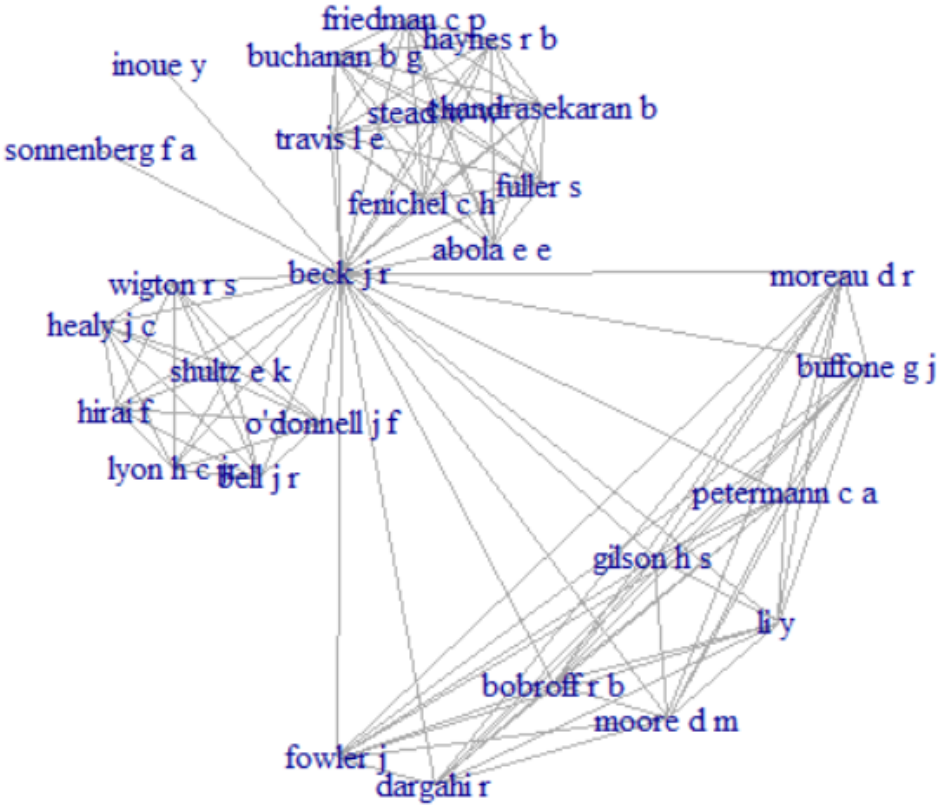


Figura 32 – Rede egocêntrica do pesquisador J. Robert Beck no período de 1991 a 1995.

4.3.2 Redes de Coautorias 1996-2000

No período de 1996 a 2000, o componente gigante concentrou a maioria dos cinco autores com maior grau, grau ponderado e intermediação. A única exceção foi o pesquisador Sergio Cerutti, identificado neste trabalho como “Cerutti S” que não pertence a este componente.

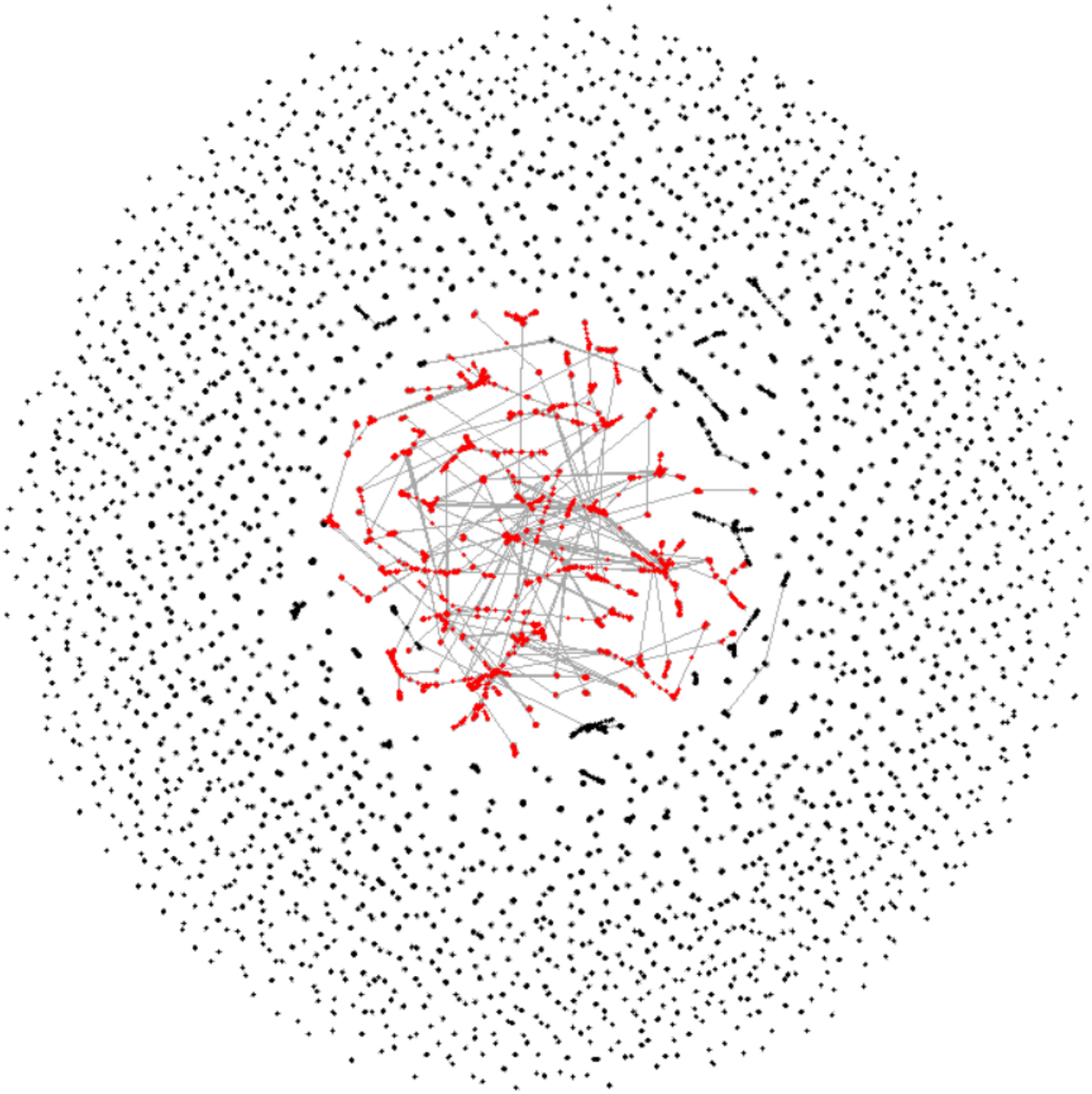


Figura 33– Representação gráfica da rede de 1996-2000 com 15.782 autores e 35.094 coautorias. A figura em alta resolução está disponível em https://bit.ly/3o9nVMT .

O autor com maior grau neste período foi o pesquisador Reinhold Haux, identificado neste trabalho como “Haux R”, com grau 74, ou seja, possui coautorias com outros 74 autores. Cabe salientar que ele foi o autor com maior centralidade de intermediação corroborando a sua importância na rede de coautoria. Também para ilustrar sua importância, o pesquisador Reinhold Haux foi eleito membro “fellow” da AMIA em 1999 (https://www.amia.org/about-amia/leadership/acmi-fellow/reinhold-haux-phd-facmi). Na Figura 34 é possível observar sua rede egocêntrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores.

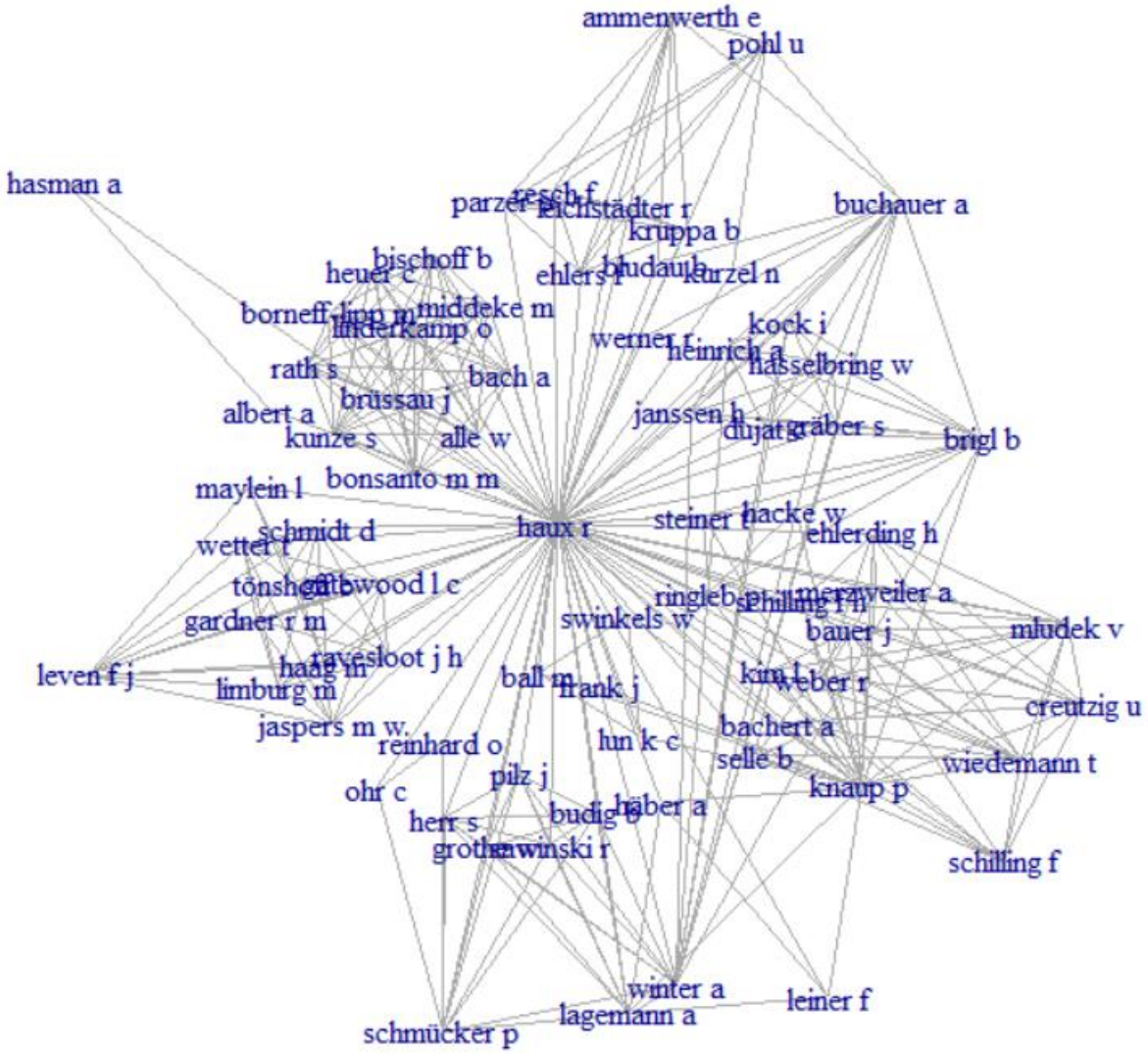


Figura 34 – Rede egocêntrica do pesquisador Renhold Haux no período de 1996 a 2000.

4.3.3 Redes de Coautorias 2001-2005

Para o período de 2001 a 2005, o componente gigante concentrou a maioria dos cinco autores com maior grau, grau ponderado e intermediação. Os autores Reinhold Haux, Fernando Martin Sanchez e Elske Ammenwerth foram identificados no segundo maior componente da rede. Esse fenômeno pode ter ocorrido devido ao componente gigante concentrar apenas 13% dos autores, um número baixo para uma rede de coautoria (6) com 30.260 coautores. Foi observado que o segundo maior componente concentra 6,49% dos coautores.

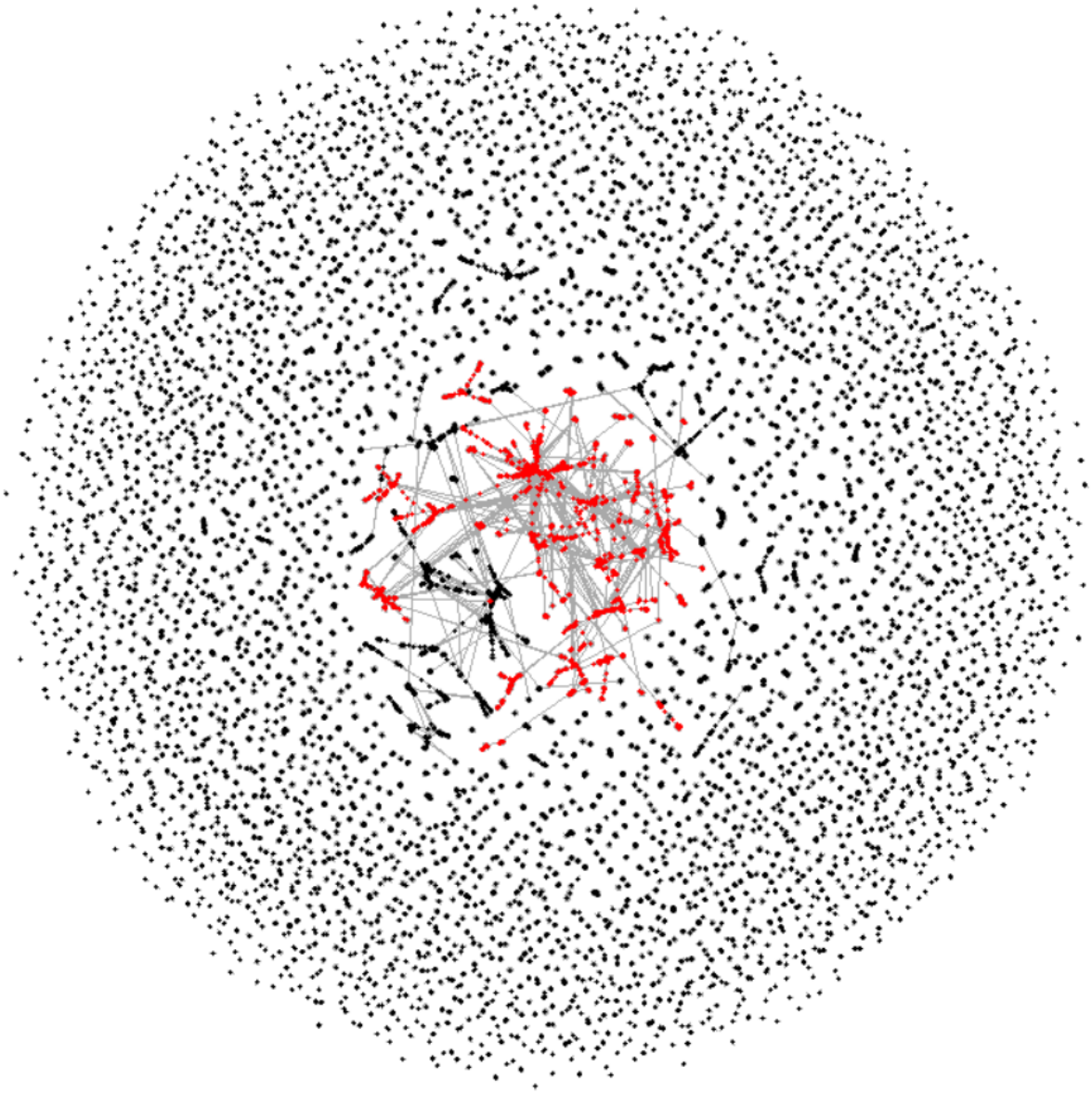


Figura 35 – Representação gráfica da rede de 2001-2005 com 31.257 autores e 78.253 coautorias. A figura em alta resolução está disponível em https://bit.ly/37l8ri5 .

O autor com maior grau neste período foi o pesquisador David W. Bates, identificado neste trabalho como “Bates David W”, com grau 135, ou seja, possui coautorias com outros 135 autores. Para ilustrar sua importância, o pesquisador David W. Bates foi eleito mebro “fellow” da AMIA em 2000 (https://www.amia.org/about-amia/leadership/acmi-fellow/david-w-bates-md-msc-facmi). Na Figura 36 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores. Devido ao número de ligações, para fins estéticos, foram omitidos os nomes dos coautores na figura.

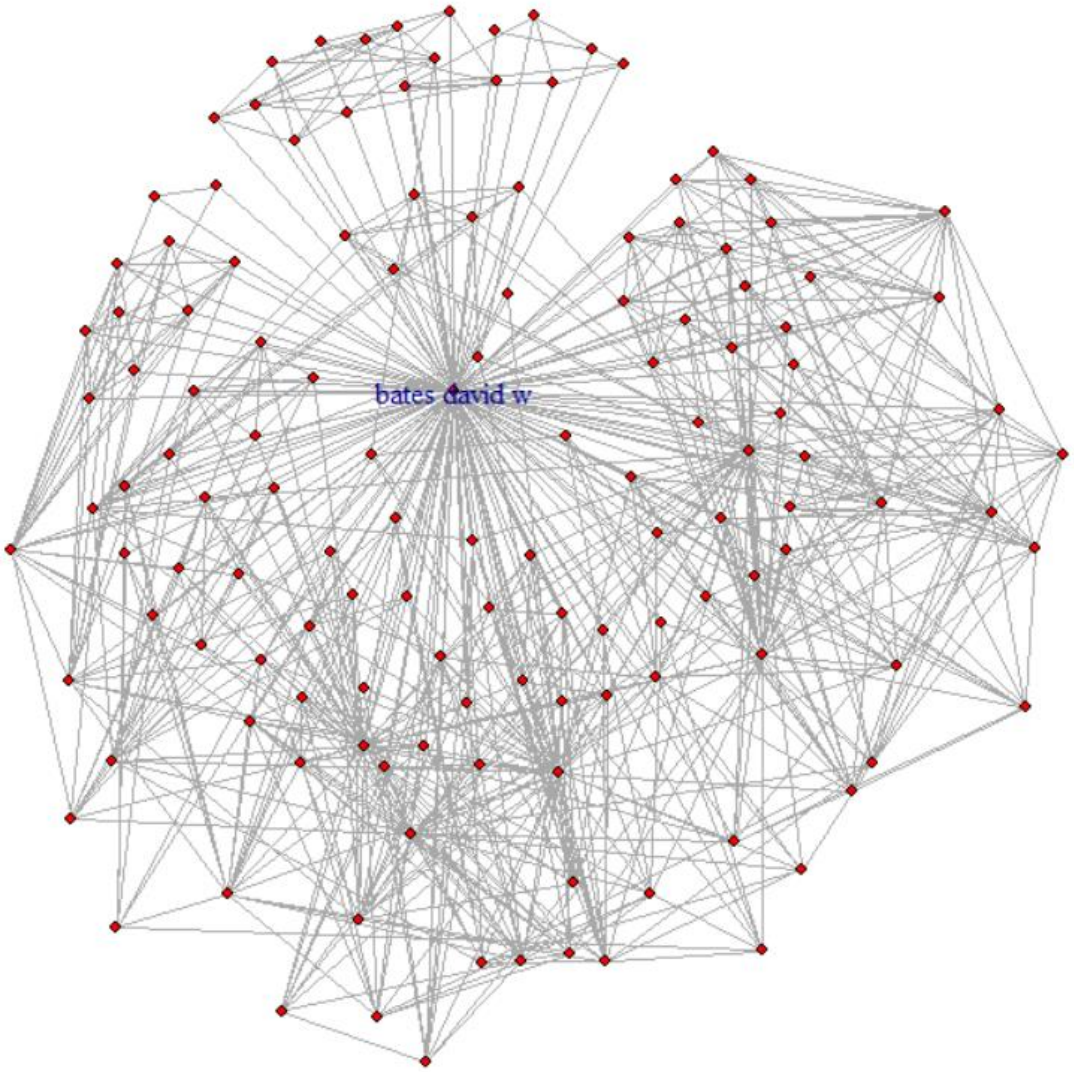


Figura 36 – Rede egocêntrica do pesquisador David W. Bates no período de 2001 a 2005.

Em relação a centralidade de intermediação, o pesquisador J. Marc Overhage, neste trabalho identificado como “Overhage J Marc”, é o autor mais central em relação a intermediação com 0,21% dos caminhos entre outros autores da rede passando por ele. Também para ilustrar sua importância, o pesquisador J. Marc Overhage foi eleito mebro “fellow” da AMIA (American Medical Informatics Association) em 1997 (https://www.amia.org/about-amia/leadership/acmi-fellow/j-marc-overhage-md-phd-facmi). Na Figura 37 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é

possível observar as coautorias entre os seus coautores. Devido ao número de ligações, para fins estéticos, foram omitidos os nomes dos coautores na figura.

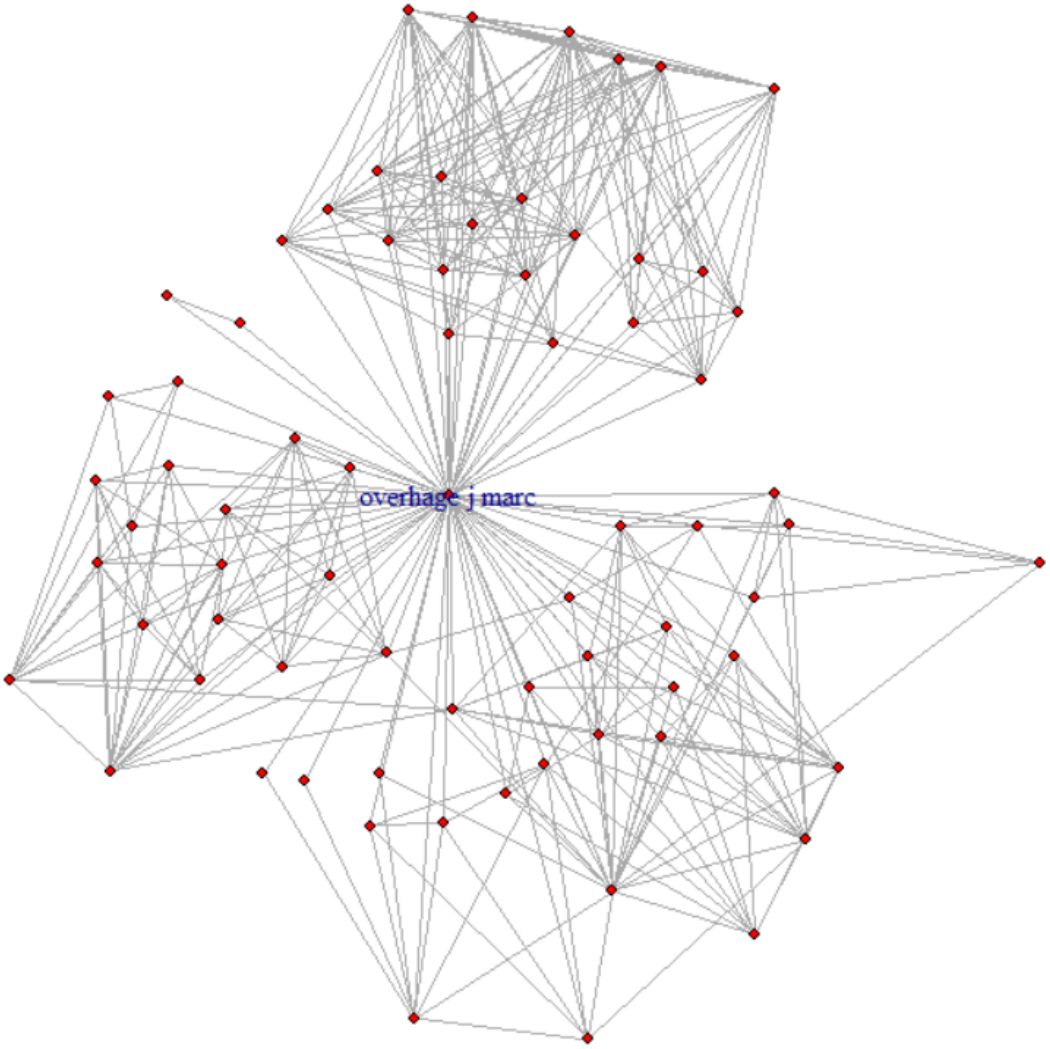


Figura 37 – Rede egocêntrica do pesquisador J. Marc Overhage no período de 2001 a 2005.

4.3.4 Redes de Coautorias 2006-2010

Para o período de 2006 a 2010, o componente gigante concentrou todos os cinco autores com maior grau, grau ponderado e intermediação.

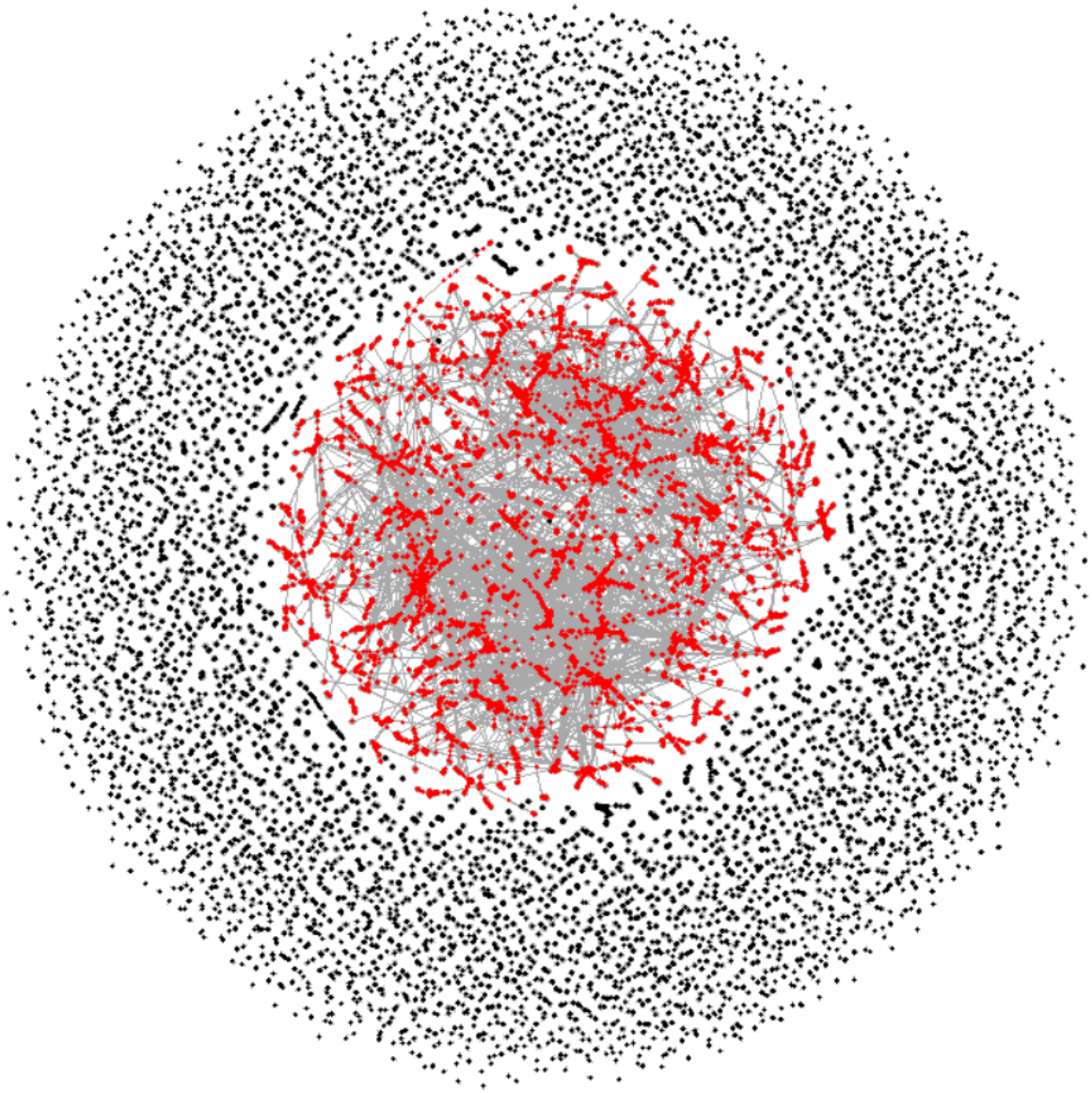


Figura 38 – Representação gráfica da rede de 2006-2010 com 55.883 autores e 162.212 coautorias. A figura em alta resolução está disponível em https://bit.ly/33uKYKk .

O autor com maior grau neste período foi novamente o pesquisador David W. Bates, neste período, com grau 176, ou seja, possui coautorias com outros 176 autores. Na Figura 39 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores. Devido ao número de ligações, para fins estéticos, foram omitidos os nomes dos coautores na figura.

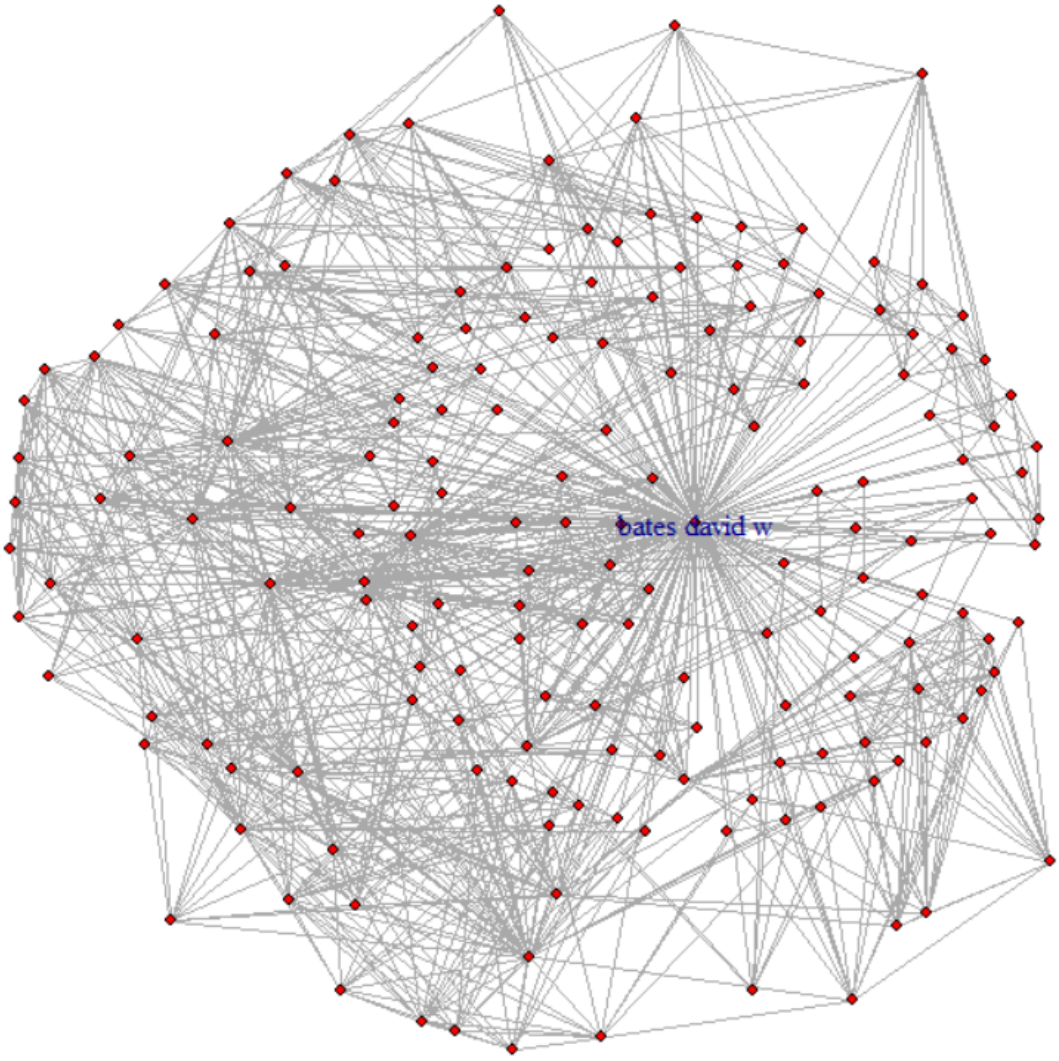


Figura 39 - Rede egocêntrica do pesquisador David W. Bates no período de 2006 a 2010.

Em relação a centralidade de intermediação, o pesquisador Xavier Pennec, neste trabalho identificado como “Pennec Xavier”, é o autor mais central em relação a intermediação com 0,74% dos caminhos entre outros autores da rede passando por ele. Também para ilustrar sua importância, o pesquisador Xavier Pennec possui mais de 200 publicações com mais de 8900 citações no total (https://publons.com/researcher/2598647/xavier-pennec/). Na Figura 40 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre os seus coautores. Devido ao número de ligações, para fins estéticos, foram omitidos os nomes dos coautores na figura.

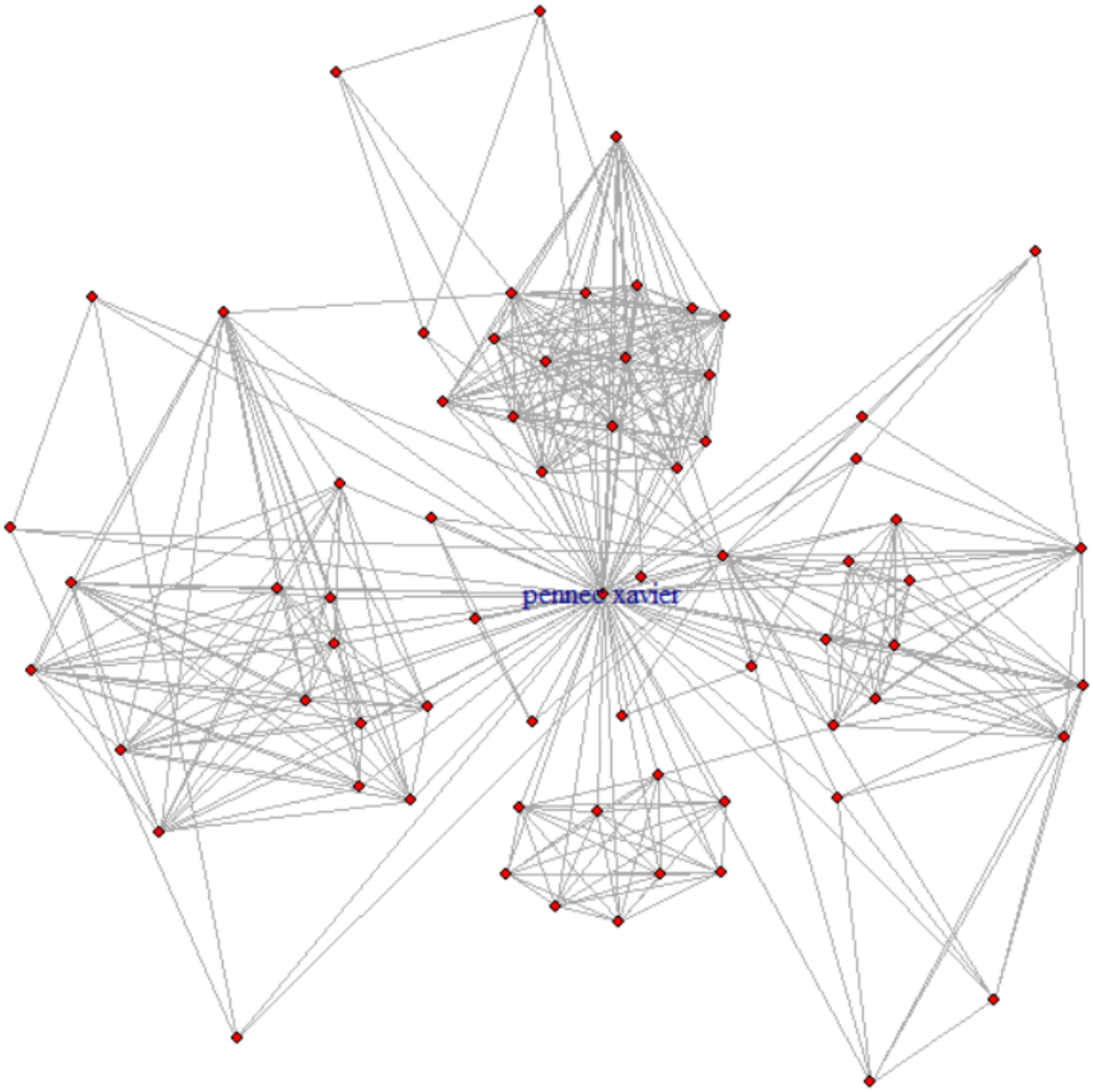


Figura 40 - Rede egocêntrica do pesquisador Xavier Pennec no período de 2006 a 2010.

4.3.5 Redes de Coautorias 2011-2015

Para o período de 2011 a 2015, o componente gigante concentrou todos os cinco autores com maior grau, grau ponderado e intermediação.

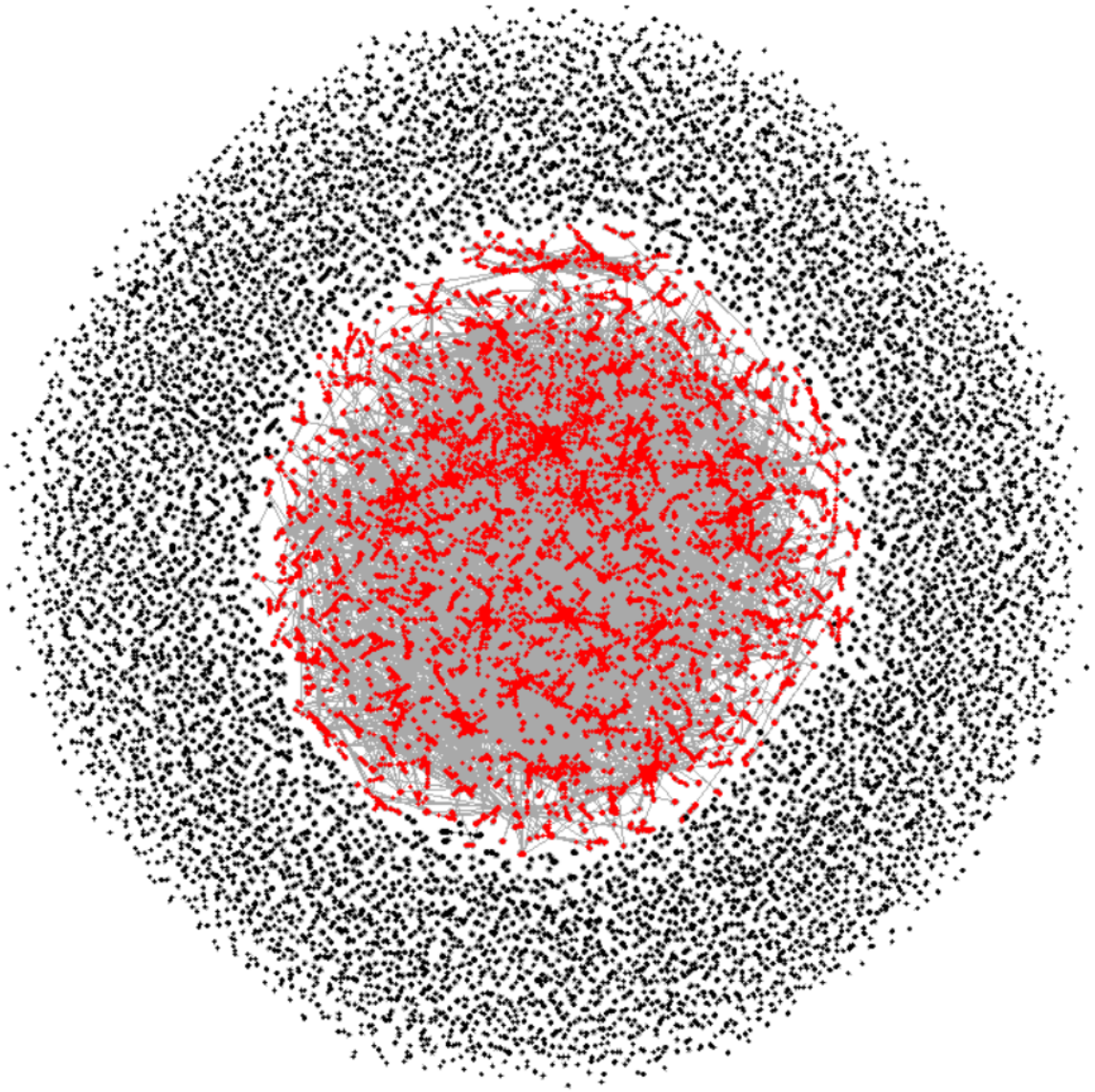


Figura 41 – Representação gráfica da rede de 2011-2015 com 92.355 autores e 342.907 coautorias. A figura em alta resolução está disponível em https://bit.ly/3qpyZrk .

O autor com maior grau identificado neste período foi Jun Wang, com grau 342. Porém, foi percebido que o nome Jun Wang pertence a diversos pesquisadores diferentes. O tratamento de desambiguação, homonimía não foi considerado neste trabalho conforme salientado no capítulo 3 – Materiais e Métodos. Assim foi considerado o segundo autor com maior grau, novamente o pesquisador David W. Bates com grau 237, ou seja, possui coautorias com outros 237 autores. O pesquisador David W. Bates também é o autor que possui a maior centralidade de intermediação, com 0,71% dos caminhos entre outros autores da rede passando por ele. Na Figura 42 é possível observar sua rede egocênctrica de ordem 1,5 composta por ele e os coautores identificados no período. Também é possível observar as coautorias entre

os seus coautores. Devido ao número de ligações, para fins estéticos, foram omitidos os nomes dos coautores na figura.

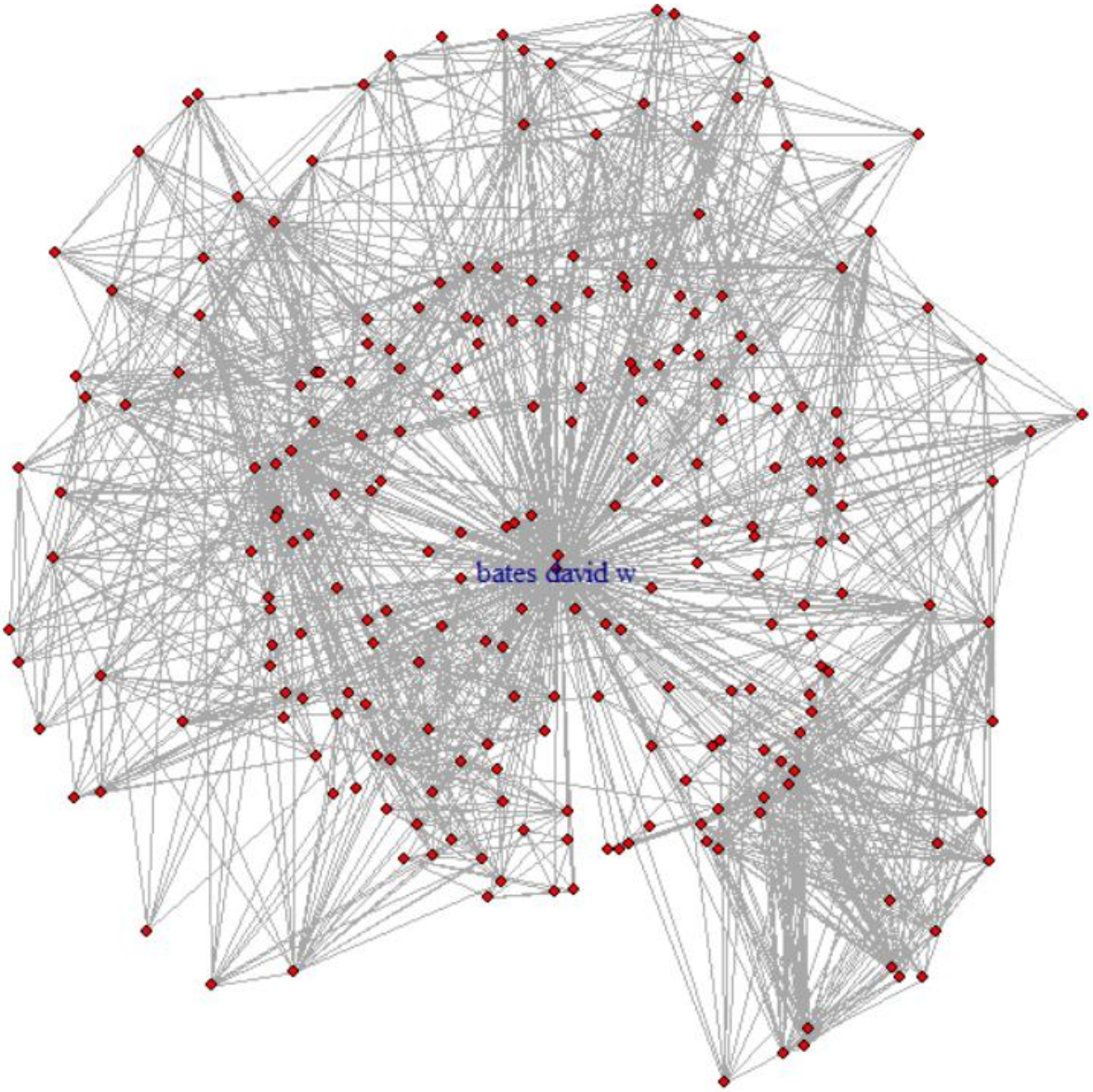


Figura 42 – Rede egocêntrica do pesquisador David W. Bates e seus coautores no período de 2011 a 2015.

5 DISCUSSÃO

Um número crescente de publicações explorou avaliar a estrutura e evolução da área de Informática em Saúde (1–5). Além das publicações realizadas pelo próprio grupo Saúde 360º, não foram encontradas publicações sobre estrutura ou evolução da área de IS utilizando-se análise de redes de coautoria ou modelagem de tópicos. Este foi um grande motivador para a realização desta pesquisa.

5.1 Etapa 1: Coleta de Artigos do PubMed

A estratégia de busca proposta neste estudo pode ser considerada original, ao selecionar apenas periódicos que tenham IS entre as palavras chaves de seus escopos, diferentemente de outros trabalhos que buscam diretamente as palavras chaves sobre IS nos artigos (4) (5).

O crescimento anual do número de artigos em IS encontrado neste trabalho pode ser considerado semelhante ao encontrado por Deshazo e colaboradores (4) se considerarmos o mesmo recorte do período entre os anos de 1991 e 2006. Comparada com a taxa de crescimento média do número de artigos total do PubMed, a taxa de crescimento média do número de artigos em IS encontrada neste trabalho é aproximadamente 3 vezes superior para o mesmo período.

No trabalho de Nadri e colaboradores (54), que buscou apresentar os 100 artigos de IS mais citados, dos dez periódicos em que foram publicados estes artigos, sete também foram identificados neste trabalho. Cabe salientar que o periódico que reúne o maior número de artigos dentre os 100 mais citados (n=71), é o periódico Statistics in Medicine, que apesar de indexado no PubMed, não possui palavras-chaves relacionadas à IS em seu escopo e por isso não foi considerado nesta pesquisa.

5.2 Modelagem Autor-Tópico

O número de 50 tópicos (k=50) foi escolhido por ser um número bem acima do número de categorias propostas existentes para a área de Informática em Saúde. O objetivo deste trabalho foi realizar uma análise exploratória e não realizar uma

adequação a categorização existente. Para fins de exemplo, o número de seções do anuário de IS da IMIA (*International Medical Informatics Association*) em 2015 (“Yearbook of Medical Informatics”, 2015) apresentava somente 11 seções (Quadro 3):

Quadro 3 – Seções do anuário da IMIA em 2015

Seção	**Descrição**
Section 1	Health and Clinical Management
Section 2	Human Factors and Organizational Issues
Section 3	Clinical Information Systems
Section 4	Sensor, Signal and Imaging Informatics
Section 5	Decision Support
Section 6	Knowledge Representation and Management
Section 7	Education and Consumer Health Information
Section 8	Bioinformatics and Translational Informatics
Section 9	Clinical Research Informatics
Section 10	Natural Language Processing
Section 11	Public Health and Epidemiology Informatics

Os resultados da modelagem autor-tópico revelaram um número maior de tópicos na área de Informática em Saúde. Na rotulação de tópicos foi possível identificar entre 68% e 88% dos 50 tópicos obtidos para cada período. Na modelagem de tópicos é comum encontrarmos tópicos em que a combinação de suas palavras mais frequentes não seja possível de se induzir um significado. Outro achado comum foi a identificação de tópicos com significados semelhantes e que poderiam ser agrupados num mesmo tópico.

Os modelos autor-tópico obtidos possuem um grande potencial, para contribuir tanto para futuras pesquisas quanto para apoio no ensino da história da IS. Apesar de não ter sido considerado uma análise de co-citação nem o número de citações para os artigos aqui identificados, os modelos conseguem trazer informações relevantes para a área de IS.

Cabe salientar que não foram encontradas referências na literatura sobre a aplicação da modelagem autor-tópico com base em artigos científicos sobre informática em saúde.

5.3 Análise de Redes de Coautorias

As redes de coautoria obtidas para os cinco períodos estudados evidenciaram por meio da caracterização feita pelas métricas globais, que se tratam de redes esparsas com pouca densidade, o que é comum para o contexto de redes de coautoria. Os resultados mostram que a densidade diminui a cada período, o que indica ser tanto pelo aumento do número de artigos ao longo dos anos, quanto pelo aumento do número de periódicos.

Os componentes gigantes de cada período concentraram praticamente todos os autores de maior relevância para a área de IS. Este fenômeno é comum conforme já relatado na literatura (9).

Cabe salientar que para o período de 2001 a 2005, houve uma redução em percentual do tamanho do componente gigante, não demonstrando a tendência de crescimento dos outros períodos. Para este período o componente gigante concentrou 3939 autores enquanto o segundo maior componente concentrou 1963 autores. Isso se deveu ao fato de que um conjunto de ligações deixou de ser identificado e o componente gigante foi dissolvido em componentes menores. A origem deste fenômeno foi uma mudança na identificação de nomes de autores no PubMed. Foi identificado por meio dos dados coletados, que a partir do ano de 2001, para alguns periódicos os nomes começaram a ser identificados com o formato sobrenome seguido do primeiro nome e nomes do meio. Antes de 2001 os autores eram identificados pelo sobrenome seguidos apenas das iniciais do primeiro nome e nomes do meio. A partir de 2002 em diante o novo formato foi progressivamente sendo adotado para todos as publicações. Assim vários importantes pesquisadores do período de 2001 a 2005 foram identificados duas vezes, uma vez com a identificação anterior e uma vez com a nova identificação em diferentes componentes da rede. Este fato demonstra a importância do tratamento da desambiguação de nomes de autores que foi uma limitação deste estudo.

Os resultados demonstraram que a maioria dos cinco autores com maior grau e centralidade de intermediação identificados nos cinco períodos são pesquisadores de alta relevância que colaboraram muito para o crescimento da área de IS.

5.4 Contribuições tecnológicas

- PubMed2DB

Oliver e colaboradores (56) já haviam apontados os benefícios de se ter uma cópia local e com uma estrutura que facilitasse as consultas de pesquisadores. Porém os autores consideraram a versão do DTD do Pubmed de 2003, a qual sofreu mais de 10 atualizaçoes até a data desta pesquisa. E apesar das rotinas criadas por eles estarem disponibilizadas publicamente, as mesmas se tornaram obsoletas, e requerem um grau de esforço considerável para a adequação a estrutura atual e futuras do DTD do PubMed. Ao utilizar um método que considera a geração automática de código-fonte para a criação e carga de banco de dados em suas etapas mais onerosas, o PubMed2DB minimiza o impacto causado por atualizações da estrutura do DTD do PubMed.

- TM.NET

Até o início deste trabalho nenhuma rotina computacional, programa, ou software que implementasse o modelo autor-tópico foi encontrado na literatura. Variações do LDA como Supervised LDA e Hierarchical LDA podem ser facilmente encontrados como em (http://www.cs.columbia.edu/~blei/topicmodeling_software.html).

- AuthorTopicViewer

Também não foi encontrado na literatura um visualizador para modelos autor-tópico, o que se fez necessário para facilitar a análise e navegação nos modelos criados. Apenas visualizadores de modelos LDA foram encontrados na literatura, sendo o LDAvis um dos mais utilizados (57).

5.5 Trabalhos Futuros

A interseção entre os modelos autor-tópico e as redes de coautoria obtidas foi um dos objetivos iniciais deste trabalho, porém evidenciou-se um esforço e complexidade bem superior ao planejado. Uma pesquisa específica com este objetivo pode utilizar as contribuições do trabalho realizado aqui e contribuir originalmente para a comunidade de IS. Os resultados deste trabalho indicam claramente uma complementaridade da aplicação das técnicas. Como exemplo podemos observar que o pesquisador Richard Wooton, que na rede de coautoria do período de 1996 a 2000 é um dos cinco autores com maior grau, ou seja um dos que mais colaboraram com

outros autores no período. Já sob o olhar do modelo autor-tópico do mesmo período podemos observar que ele é identificado como um dos mais relevantes para o tópico sobre Telemedicina.

Outra possibilidade é uma avaliação do foco dos autores por meio da avaliação da dispersão da distribuição de probabilidades de tópicos dos autores.

A partir dos resultados desta tese também é possível uma análise sob o ponto de vista regional. Será que pesquisadores brasileiros tem interesses científicos mais concentrados em determinados tópicos?

6 CONCLUSÃO

Este trabalho buscou apresentar um mapa de interesses de autores em informática em saúde por meio da aplicação de modelagem autor-tópico e uma análise de redes de coautoria sobre os mesmos autores e períodos.

Os modelos autor-tópico obtidos apresentaram resultados consistentes, servindo como uma alternativa para entender melhor a evolução da área de IS do ponto de vista dos interesses dos autores identificados pelos tópicos obtidos.

A análise de redes de coautoria evidenciou a evolução da estrutura de colaboração global ao longo dos anos, assim como uma visão local da importância dos autores por meio de métricas de centralidade.

Os resultados obtidos neste trabalho são bem amplos pois a ideia foi realizar uma análise exploratória sob o olhar das técnicas de modelagem autor-tópico e de análise de redes de coautorias. Os resultados podem servir de base para estudos mais profundos e específicos acerca do tema.

As contribuições tecnológicas aqui desenvolvidas podem facilitar novos estudos tanto da literatura de Informática em Saúde quanto de outras áreas sejam da Saúde ou não, assim como em outras fontes de dados que não somente o PubMed.

REFERÊNCIAS

1. Demiris G. Interdisciplinary innovations in biomedical and health informatics graduate education. Methods Inf Med. 2007;46(1):63–6.

2. Morris TA, McCain KW. The Structure of Medical Informatics Journal Literature. J Am Med Inform Assoc. 1998;5(5):448–66.

3. Greenes RA, Siegel ER. Characterization of an Emerging Field: Approaches to Defining the Literature and Disciplinary Boundaries of Medical Informatics. Proc Annu Symp Comput Appl Med Care. 4 de novembro de 1987;411–5.

4. DeShazo JP, LaVallie DL, Wolf FM. Publication trends in the medical informatics literature: 20 years of “Medical Informatics” in MeSH. BMC Medical Informatics and Decision Making. 21 de janeiro de 2009;9:7.

5. Lyu P-H, Yao Q, Mao J, Zhang S-J. Emerging medical informatics research trends detection based on MeSH terms. Informatics for Health and Social Care. 3 de julho de 2015;40(3):210–28.

6. Sameer Kumar. Co-authorship networks: a review of the literature. Aslib Journal of Info Mgmt. 14 de janeiro de 2015;67(1):55–73.

7. Borgatti SP, Mehra A, Brass DJ, Labianca G. Network Analysis in the Social Sciences. Science. 13 de fevereiro de 2009;323(5916):892–5.

8. Girvan M, Newman MEJ. Community structure in social and biological networks. PNAS. 6 de novembro de 2002;99(12):7821–6.

9. Otte E, Rousseau R. Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science. 12 de janeiro de 2002;28(6):441–53.

10. Newman MEJ. The structure of scientific collaboration networks. PNAS. 16 de janeiro de 2001;98(2):404–9.

11. Freire VPM, Figueiredo DR. Ranking in Collaboration Networks Using a Relationship Intensity Metric. In: 2010 Brazilian Symposium of Collaborative Systems - Simposio Brasileiro de Sistemas Colaborativos (SBSC). 2010. p. 71–8.

12. Wu Y, Duan Z. Social network analysis of international scientific collaboration on psychiatry research. Int J Ment Health Syst. 2015;9(1):2.

13. Vacca R, McCarty C, Conlon M, Nelson DR. Designing a CTSA-Based Social Network Intervention to Foster Cross-Disciplinary Team Science. Clin Transl Sci. 19 de março de 2015;

14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. março de 2003;3:993–1022.

15. Buntine W, Jakulin A. Applying Discrete PCA in Data Analysis. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence [Internet]. Arlington, Virginia, United States: AUAI Press; 2004 [citado 5 de abril de 2015]. p. 59–66. (UAI '04). Disponível em: http://dl.acm.org/citation.cfm?id=1036843.1036851

16. Blei DM, Lafferty JD. A Correlated Topic Model of Science. The Annals of Applied Statistics. 1º de junho de 2007;1(1):17–35.

17. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The Author-topic Model for Authors and Documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence [Internet]. Arlington, Virginia, United States: AUAI Press; 2004 [citado 5 de abril de 2015]. p. 487–494. (UAI '04). Disponível em: http://dl.acm.org/citation.cfm?id=1036843.1036902

18. Colepicolo E. Epistemologia da Informática em Saúde: entre a teoria e a prática [Internet] [Mestrado]. Universidade Federal de São Paulo; 2008. Disponível em: http://www.disacad.unifesp.br/sapg/arquivos/arq_55.pdf

19. Teixeira FO. Classificacao e indexacao de artigos cientificos internacionais de Informática em Saúde [Internet]. Universidade Federal de São Paulo (UNIFESP); 2011 [citado 29 de novembro de 2020]. Disponível em: https://repositorio.unifesp.br/handle/11600/21726

20. Humphrey SM. A New Approach to Automatic Indexing Using Journal Descriptors. Proceedings of the ASIS Annual Meeting. 1998;35:496–500.

21. Costa TM. LattesRank do CBIS2008: Ranking dos Participantes do Congresso Brasileiro de Informática em Saúde 2006 Baseado no Grau de Conexão via Currículo Lattes [Internet]. Poster apresentado em: CBIS 2008; 2008 mar 12 [citado 16 de agosto de 2015]; Campos de Jordão - SP. Disponível em: www.sbis.org.br/site/site.dll/view?pagina=131

22. Baptista RS, Hummel AD, Teixeira FO, Pisa IT. Scientific collaboration in Brazilian Health Informatics community. In: 1st European Conference on Social Networks Abstract Book [Internet]. Barcelona, Espanha; 2014 [citado 29 de novembro de 2020]. p. 17. Disponível em: https://jornades.uab.cat/eusn/sites/jornades.uab.cat.eusn/files/Abstract_book_eusn.pdf

23. Baptista RS, Araújo GD, Teixeira FO, Pisa IT. Scientific Collaboration in Brazilian Health Informatics Scientific Community. In: XXXVI International Sunbelt Social Network Conference Presentation and Poster Abstract [Internet]. California, EUA; [citado 29 de novembro de 2020]. p. 14–5. Disponível em: http://sunbelt2016.insna.org/wp-content/uploads/2015/09/Sunbelt2016abstracts.pdf

24. Brito TD de LV. Análise da colaboração nos Grupos de Interesse Especial (SIG) da Rede Universitária de Telemedicina (RUTE) [Internet]. Universidade Federal de São Paulo (UNIFESP); 2016 [citado 29 de novembro de 2020]. Disponível em: https://repositorio.unifesp.br/handle/11600/41275

25. Jeong S, Lee SK, Kim H-G. Knowledge Structure of Korean Medical Informatics: A Social Network Analysis of Articles in Journal and Proceedings. Healthcare Informatics Research. 1º de março de 2010;16(1):52–9.

26. Baptista RS, Brito TD de LV, Braun LL, Tenório JM, Pisa IT. Colaboração acadêmica em informática em saúde baseada em análise de redes sociais. Journal of Health Informatics [Internet]. 7 de dezembro de 2019 [citado 30 de maio de 2020];11(4). Disponível em: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/669

27. Chen Q, Ai N, Liao J, Shao X, Liu Y, Fan X. Revealing topics and their evolution in biomedical literature using Bio-DTM: a case study of ginseng. Chinese Medicine. 12 de setembro de 2017;12(1):27.

28. Wang S-H, Ding Y, Zhao W, Huang Y-H, Perkins R, Zou W, et al. Text mining for identifying topics in the literatures about adolescent substance use and depression. BMC Public Health. 19 de março de 2016;16(1):279.

29. Kongthon A, Haruechaiyasak C, Thaiprayoon S. Enhancing the Literature Review Using Author-Topic Profiling. In: Buchanan G, Masoodian M, Cunningham SJ, organizadores. Digital Libraries: Universal and Ubiquitous Access to Information. Berlin, Heidelberg: Springer; 2008. p. 335–8. (Lecture Notes in Computer Science).

30. Kusumawardani RP, Khairunnisa SO. Author-Topic Modelling for Reviewer Assignment of Scientific Papers in Bahasa Indonesia. In: 2018 International Conference on Asian Language Processing (IALP). 2018. p. 351–6.

31. Anand CV and H. Research Methodology. 1st edition. PEARSON INDIA; 2017.

32. Wainer J. Métodos de Pesquisa Quantitativa e Qualitativas para Ciência da Computação. UNICAMP, SP, http://www ic unicamp br/~ wainer/papers/metod07 pdf. 2007;

33. Garcia CC, Martrucelli CRN, Rossilho M de MF, Denardin OVP. Autoria em artigos científicos: os novos desafios. Brazilian Journal of Cardiovascular Surgery. dezembro de 2010;25(4):559–67.

34. Hilário CM, Grácio MCC, Guimarães JAC. Aspectos éticos da coautoria em publicações científicas. Em Questão. 19 de abril de 2018;24(2):12–36.

35. Bufrem LS, Gabriel Junior RF, Gonçalves V. Práticas de co-autoria no processo de comunicação científica na pós-graduação em Ciência da Informação no Brasil. Informação & Informação. 15 de dezembro de 2010;15(1esp):111.

36. Promoting integrity in scholarly research and its publication | COPE: Committee on Publication Ethics [Internet]. [citado 29 de novembro de 2020]. Disponível em: https://publicationethics.org/

38. Mcauliffe J, Blei D. Supervised Topic Models. Advances in Neural Information Processing Systems. 2007;20:121–8.

39. Estrada E, Knight PA. A first course in network theory. Oxford University Press, USA; 2015.

40. Easley D, Kleinberg J, others. Networks, crowds, and markets. Vol. 8. Cambridge university press Cambridge; 2010.

41. PubMed. In: Wikipedia [Internet]. 2020 [citado 15 de novembro de 2020]. Disponível em: https://en.wikipedia.org/w/index.php?title=PubMed&oldid=985836229

42. Entrez Programming Utilities Help. National Center for Biotechnology Information (US); 2010.

43. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. NLM Catalog [Internet]. NLM Catalog Help [Internet]. National Center for Biotechnology Information (US); 2019 [citado 10 de outubro de 2020]. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK3799/

44. Feinerer I, Hornik K, Software A, Ghostscript) I (pdf_info ps taken from G. tm: Text Mining Package [Internet]. 2019 [citado 11 de outubro de 2020]. Disponível em: https://CRAN.R-project.org/package=tm

45. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data. 19 de outubro de 2019;6(1):93.

46. Hussain I, Asghar S. A survey of author name disambiguation techniques: 2010–2016. The Knowledge Engineering Review [Internet]. ed de 2017 [citado 29 de novembro de 2020];32. Disponível em: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/abs/survey-of-author-name-disambiguation-techniques-20102016/EF8B67C2D3BFABBB05F883C899A9934A

47. Kang I-S, Na S-H, Lee S, Jung H, Kim P, Sung W-K, et al. On co-authorship for author disambiguation. Information Processing & Management. 1º de janeiro de 2009;45(1):84–97.

48. Strotmann A, Zhao D, Bubela T. Author name disambiguation for collaboration network analysis and visualization. Proceedings of the American Society for Information Science and Technology. 2009;46(1):1–20.

49. Berners-Lee TJ. The world-wide web. Computer Networks and ISDN Systems. 1º de novembro de 1992;25(4):454–9.

50. Journal Indexing and Metrics: Journal of Telemedicine and Telecare [Internet]. SAGE Journals. [citado 13 de julho de 2020]. Disponível em: https://journals.sagepub.com/metrics/jtt

51. Begam BF, Kumar JS. A Study on Cheminformatics and its Applications on Modern Drug Discovery. Procedia Engineering. 1º de janeiro de 2012;38:1264–75.

52. Arnaboldi V, Conti M, Passarella A, Pezzoni F. Analysis of Ego Network Structure in Online Social Networks. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. 2012. p. 31–40.

53. ACMI Fellowship | AMIA [Internet]. [citado 21 de novembro de 2020]. Disponível em: https://www.amia.org/acmi-fellowship

54. Nadri H, Rahimi B, Timpka T, Sedghi S. The Top 100 Articles in the Medical Informatics: a Bibliometric Analysis. J Med Syst. 1º de outubro de 2017;41(10):150.

55. Thieme E-Journals - Yearbook of Medical Informatics / Abstract [Internet]. [citado 4 de julho de 2020]. Disponível em: https://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0038-1639038

56. Oliver DE, Bhalotia G, Schwartz AS, Altman RB, Hearst MA. Tools for loading MEDLINE into a local relational database. BMC Bioinformatics. 7 de outubro de 2004;5(1):146.

57. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces [Internet]. Baltimore, Maryland, USA: Association for Computational Linguistics; 2014 [citado 31 de março de 2020]. p. 63–70. Disponível em: http://aclweb.org/anthology/W14-3110

ANEXOS

Anexo 1 – Aprovação do Comitê de Ética em Pesquisa – UNIFESP – HSP

UNIVERSIDADE FEDERAL DE SÃO PAULO HOSPITAL SÃO PAULO UNIFESP-HSP

Continuação do Parecer: 1.156.458

científica em Ciências da Saúde baseada em coautorias de artigos científicos. Objetivo 2: Propor modelo autor-tópico (AT) para categorizar interesses acadêmi-cos em ciências da saúde baseado em resumos e autorias de artigos científicos. Objetivo 3: Integrar análise de redes sociais (SNA) e modelo autor-tópico (AT) pa-ra aprofundar a análise exploratória de uma rede de colaboração cien-tífica em ciências da saúde. Objetivo 4: Aplicar análise de redes sociais (SNA) integrado ao modelo AT pro-posto como estudo de caso de uma rede de colaboração científica em informática em saúde.

Avaliação dos Riscos e Benefícios:
Conforme parecer n: 1.113.180, de 18/6/2015

Comentários e Considerações sobre a Pesquisa:
Conforme parecer n: 1.113.180, de 18/6/2015

Considerações sobre os Termos de apresentação obrigatória:
Conforme parecer n: 1.113.180, de 18/6/2015

Recomendações:
Conforme parecer n: 1.113.180, de 18/6/2015

Conclusões ou Pendências e Lista de Inadequações:
A pendência anteriormente apontada (ver abaixo) foi atendida. ESTUDO APROVADO .

1- Na metodologia, foi informado que a etapa (objetivo 4) será baseada na aplicação de um questionário a um conjunto de pesquisadores identificados nos tópicos com a finalidade de validar se o pesquisador entende o tópico como um de seus interesses acadêmicos e se reconhece outros pesquisadores no tópico. Desta forma, será necessário dar mais detalhes sobre onde e como será aplicado este questionário e será necessário aplicar o Termo de Consentimento Livre e Esclarecido (TCLE) a estes pesquisadores. Favor enviar o modelo de TCLE. (ver modelo de TCLE na página do CEP/UNIFESP, Plataforma Brasil: "evite pendências") e enviar o modelo do questionário.
RESPOSTA/ANÁLISE CEP: foi enviado modelo de TCLE (para aplicação por via eletrônica). PENDÊNCIA ATENDIDA

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

Endereço: Rua Botucatu, 572 1º Andar Conj. 14
Bairro: VILA CLEMENTINO **CEP:** 04.023-061
UF: SP **Município:** SAO PAULO
Telefone: (11)5571-1062 **Fax:** (11)5539-7162 **E-mail:** secretaria.cepunifesp@gmail.com

Anexo 2 – PubMed2DB

Apesar de existirem padrões e aplicações para manipulação de arquivos XML, a criação de rotinas computacionais para manipular e extrair as informações desejadas de um conjunto de arquivos XML pode ser um trabalho muito árduo. Uma das abordages para facilitar a manipulação do conteúdo de arquivos XML, é a conversão de sua estrutura e conteúdo para uma estrutura que seja de manipulação mais fácil. Neste trabalho foi construída uma aplicação computacional para extração do conteúdo dos arquivos XML obtidos no passo anterior para uma estrutura de banco de dados relacional, o que facilita a manipulação e consulta por meio da linguagem SQL. Esta escolha se deveu a longa experiência do pesquisador em modelagem de dados, linguagem SQL e administração de servidores de banco de dados relacionais. Esta escolha também foi motivada porque a linguagem SQL é uma linguagem manipulação de dados muito difundida na indústria de software.

Os artigos do PubMed seguem uma estrutura padrão de dados, esta estrutura é disponibilizada por meio de um documento DTD (*Document Type Definition*). Sua versão corrente pode ser obtida no endereço https://www.nlm.nih.gov/databases/download/pubmed_medline.html. Um documento DTD representa a estrutura de elementos, atributos e regras para a criação de um documento XML e são utilizados para validar se um arquivo XML segue a estrutura definida do seu respectivo DTD. Como o DTD dos artigos do PubMed possui uma estrutura muito complexa, seria necessário um esforço de desenvolvimento muito grande modelar manualmente um banco de dados relacional que representasse a estrutura deste documento DTD, assim como criar uma rotina de extração, transformação e carga (*Extraction, Transformation and Load, ETL*) dos arquivos XML para as tabelas do banco de dados criado. Oliver e colaboradores realizaram exatamante este grande trabalho em 2004 (56). Os autores construíram e disponibilizaram estas rotinas tomando como base a versão do DTD do ano de 2003. Como o DTD das citações do PubMed passou por mais de dez revisões e atualizações desde então, a estrutura de banco de dados proposta e disponibilizada pelos autores não representa mais a estrutura corrente do DTD das citações do PubMed.

Para evitar realizar um trabalho similar que funcionasse apenas com a versão do DTD corrente, foi desenvolvida uma aplicação computacional baseada em um método de geração de código-fonte automático que pode ser executado para a ver-

são do DTD que existir no momento de sua construção. Esta abordagem possibilita uma rápida adequação da aplicação evitando que esta deixe de funcionar ou no mínimo deixe de representar a totalidade da estrutura das citações após uma ou mais atualizações do DTD das citações do PubMed.

Para o desenvolvimento do método PubMed2DB foram utilizados os seguintes recursos de software, todos eles de livre utilização:

- Visual Studio Express Edition 2012
- MS SQL Server 2016 Developer Edition
- xsd.exe
- MS Entity Framework 6

A seguir são descritos os passos necessários para o desenvolvimento do método PubMed2DB:

Passo 1: Conversão do DTD

Como primeiro passo foi necessária a conversão do documento DTD para um documento XSD (*XML Schema Definition*), que é considerado o sucessor do DTD. Apesar de ambos coexistirem, um dos benefícios do uso do XSD é ser escrito com base no próprio formato XML.

Esta conversão foi realizada por meio do editor XML encontrado no MS Visual Studio 2012. Nele foi aberto o arquivo DTD e em seguida selecionada a opção "*Create Schema*".

Passo 2: Criação do conjunto de classes

O documento XSD obtido no passo anterior foi submetido ao software xsd.exe (https://docs.microsoft.com/pt-br/dotnet/standard/serialization/xml-schema-definition-tool-xsd-exe) para converter sua estrutura em um modelo de classes implementado na linguagem de programação Microsoft C#.NET. Tendo como entrada apenas o arquivo XSD, este software criou toda a estrutura de classes necessária para representar o XSD submetido, assim como as rotinas de leitura, carga e transformação dos arquivos XML para esta estrutura de classes.

Passo 3: Mapeamento Objeto-Relacional

Foi utilizado o MS Entity Framework (EF), que é um mapeador objeto-relacional, para mapear automaticamente a estrutura de classes obtida no passo anterior à um banco de dados relacional. Este mapeamento executou a criação do banco de dados segundo a estrutura de classes definida, assim como encapsula as operações de armazenamento dos objetos instanciados pelas classes em registros em tabelas do banco de dados. Foi necessário apenas adicionar poucas linhas de código manualmente, para ligar as classes ao EF e para apontar como origem, o caminho de leitura sequencial dos arquivos XML coletados previamente e como destino, as informações para conexão ao servidor de banco de dados. O servidor de banco de dados relacional utilizado como destino foi o MS SQL Server 2016 Developer Edition.

Como saída da aplicação PubMed2DB, uma estrutura de banco de dados é criada e posteriormente populada com o conteúdo dos arquivos XML coletados. A partir do banco de dados obtido é possível realizar diversas consultas SQL para extração de dados para análises.

Anexo 3 – TM.NET

Até o início desta pesquisa, não foram encontrados softwares ou rotinas computacionais que implementassem a modelagem autor-tópico. Apenas softwares e rotinas que implementam a modelagem de tópicos por LDA e algumas de suas variações foram encontrados. Assim foi decidido construir um software para implementar a modelagem autor-tópico.

Dentre as opções encontradas que implementam o LDA, foi escolhido uma aplicação computacional de código-fonte aberto e desenvolvida na linguagem de programação C#.NET chamada de Fast Parallel Topic Model (https://olney.ai/category/2010/01/01/fastparallelcode.html). O motivo desta escolha se deu pela forma didática de sua escrita e pelo bom desempenho apresentado. Outro motivo foi a familiaridade do pesquisador na linguagem de programação em questão.

O TM.Net foi desenvolvido na linguagem C#.Net com base no .NET Framework 4.5 (https://dotnet.microsoft.com/learn/dotnet/what-is-dotnet) e com as seguintes bibliotecas de uso gratuito adicionais:

- Accord.Net
- MathNet.Numerics
- Newtonsoft.Json

O TM.Net foi desenvolvido com as seguintes funcionalidades:

Pré-processamento

A primeira parte implementada nesta funcionalidade foi a leitura de arquivos texto no formato csv. Duas rotinas de leitura foram implementadas. Uma para leitura do corpus, ou seja, os títulos e resumos, em que cada linha do arquivo representa um artigo. A segunda rotina é para leitura dos coautores de cada artigo, nesta rotina, cada linha representa os coautores de um artigo.

A segunda parte compreende o pré-processamento propriamente dito, com as rotinas de limpeza, padronização, preparação para as etapas seguintes:

- Conversão para minúsculo: esta função transforma todos os caracteres alfabéticos encontrados em caracteres minúsculos para padronização dos dados;
- Remoção de números e caracteres não alfabéticos;
- Remoção de *stopwords*: Remoção do corpus de palavras encontradas numa lista em um arquivo de texto a ser submetido pelo usuário;
- Remoção de espaços extras;
- Remoção de palavras com alta e baixa frequência no corpus: Esta rotina possui dois parâmetros, sendo estes, limite inferior e limite superior;
- Criação do conjunto de dados para processamento: foram desenvolvidos dois tipos de conjunto de dados. O primeiro para geração do modelo LDA, e é composto por um vetor de palavras únicas chamado aqui de vocabulário e um conjunto de vetores de ocorrências de palavras do vocabulário, em que cada vetor representa um artigo. O segundo tipo é para a geração do modelo autor-tópico. É composto pelo mesmo conjunto de dados anterior mais um vetor de autores únicos e um conjunto de vetores de coautores, em que cada vetor representam os autores de um artigo;

Gerador do modelo de tópicos

Apesar desta pesquisa ter como base o uso da modelagem autor-tópico, além da implementação deste modelo, um segundo modelo baseado em LDA foi desenvolvido inicialmente como aprendizado e aprofundamento do pesquisador no tema.

- Gerador do modelo LDA: Modelagem de tópicos implementado por LDA. A partir do código-fonte existente, um novo gerador foi desenvolvido, com foco em melhoria de desempenho por meio de implementações em programação paralela do .NET Framework. Para a execução do gerador, como parâmetros de entrada, além do conjunto de dados obtido após o pré-processamento, é preciso informar outros três parâmetros, sendo estes, o número de tópicos desejado e os dois hiperparâmetros alfa e beta. Como saída são geradas duas matrizes, phi e theta. Phi é uma matriz que representa a distribuição de probabilidade de palavras do vocabulário ocorrerem para cada tópico. E theta é uma

matriz que representa a distribuição de probabilidade dos tópicos ocorrerem para cada documento.

- Gerador do modelo autor-tópico. Derivado do gerador do modelo LDA criado anteriormente. Assim como no gerador do modelo LDA, são necessários como parâmetros de entrada, além do conjunto de dados obtido após o pré-processamento, neste caso incluídas as informações de coautoria, é preciso informar também, o número de tópicos desejado e os dois hiperparâmetros alfa e beta. Como saída são geradas duas matrizes, phi e theta. Phi é uma matriz que representa a distribuição de probabilidade de palavras do vocabulário ocorrerem para cada tópico. E theta é uma matriz que representa a distribuição de probabilidade dos tópicos ocorrerem para cada autor.

Exportação de resultados

- Exportação de dados para visualização e exploração de modelo LDA e modelo autor-tópico no TopicViewer
- Exportação de dados para visualização e exploração de modelo LDA e modelo autor-tópico no LDAVIS;
- Exportação de matriz de adjacência de coautorias para análise de redes de coautoria;
- Estatística descritiva do corpus (em desenvolvimento);
- Cálculo de Medidas de qualidade de tópicos (em desenvolvimento);

Anexo 4 – TopicViewer

Apesar de existirem alguns softwares para visualização, não foi encontrado nenhum software para visualização de um modelo autor-topico até o início desta pesquisa. Foi decidido então desenvolver uma aplicação web para visualização e navegação de um modelo autor-tópico.

O TopicViewer foi desenvolvido com HTML5, CSS3 e com base no framework ASP.NET MVC versão 5 e linguagem C#.Net. Também foi utilizada extensão LinqToCSV.

O TopicViewer possui apenas funcionalidades de visualização e navegação de um modelo autor-tópico submetido. Como conjunto de dados, são necessários nove arquivos texto, que representam o modelo de tópicos a serem visualizados.

Página de visualização da lista de tópicos

Nesta página, todos os tópicos são apresentados, cada tópico com seu título, uma lista com as cinco palavras com maior distribuição de frequência no tópico e uma lista com os cinco autores com maior distribuição de frequência no tópico. Ao lado de cada palavra, assim como de cada autor, é apresentada a distribuição de frequência destas no tópico. Para cada tópico há um hiperlink para a página de visualização do tópico assim como para cada autor apresentado há um hiperlink para a página de visualização do autor (Figura 1).

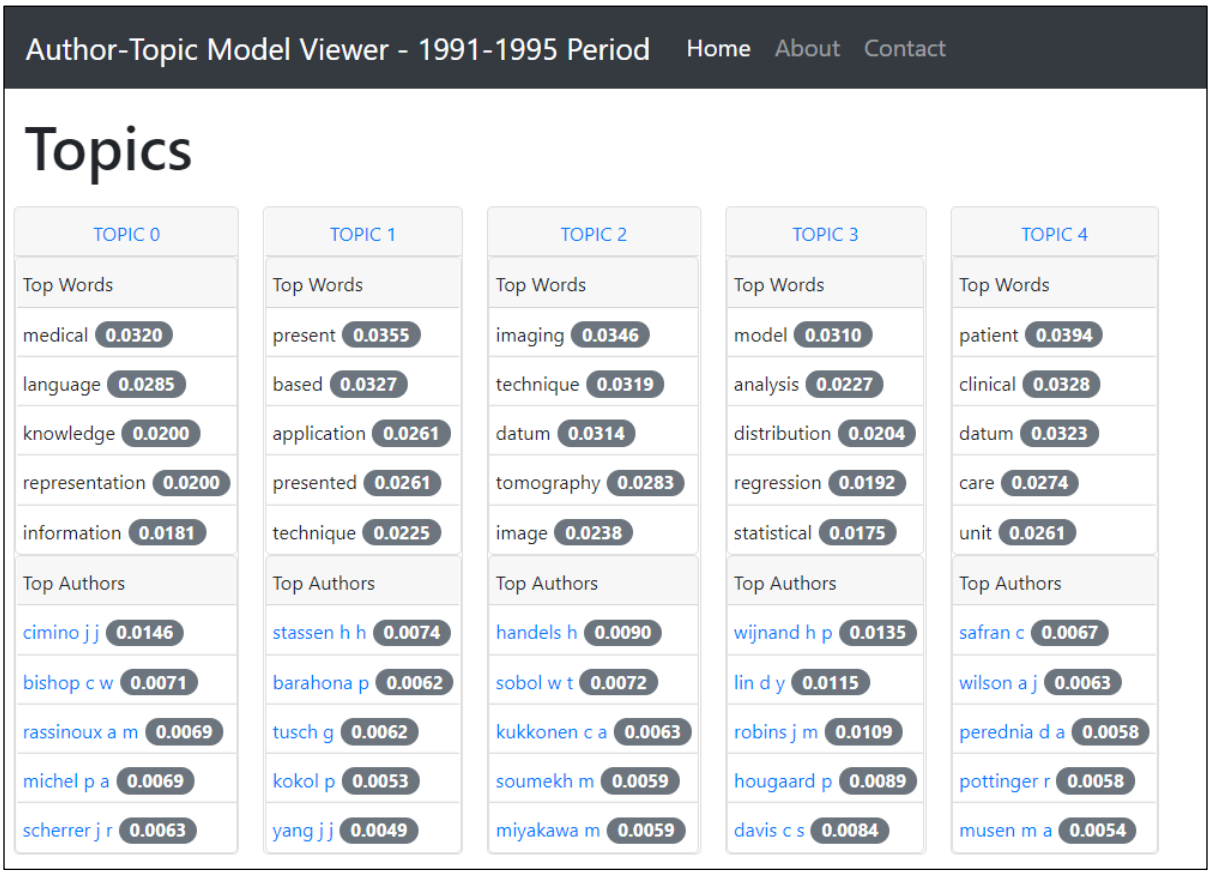


Figura 1 – Exemplo de recorte da página de visualização de tópicos

Página de visualização do tópico

Na página de visualização de tópico, pode-se ter um detalhe maior do tópico com uma lista com os 30 autores com maior distribuição de frequência para o tópico em questão e uma lista com as 30 palavras com maior distribuição no tópico. Para cada autor apresentado há um hiperlink para a página de visualização do autor (Figura 2).

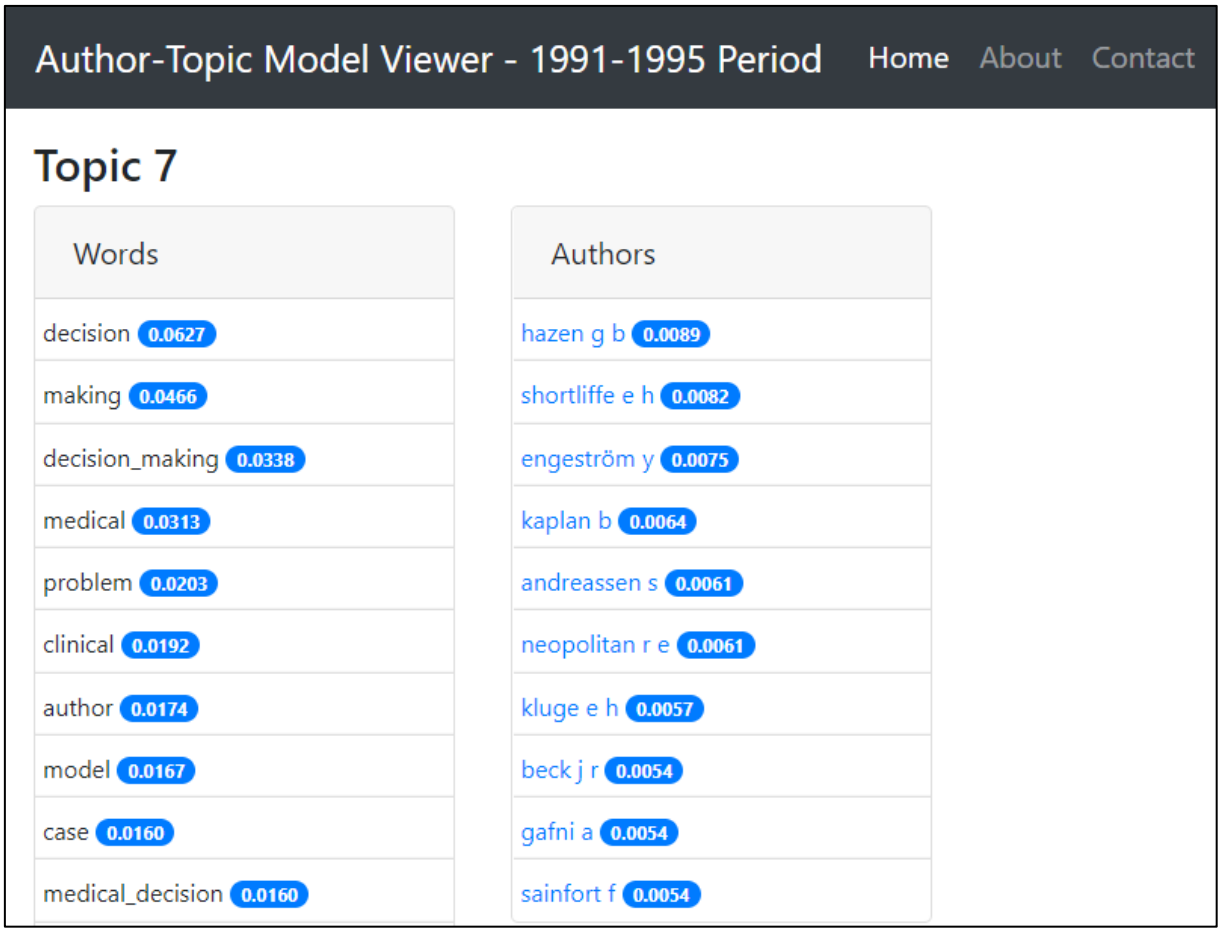


Figura 2 – Exemplo de recorte da página de visualização do tópico

Página de visualização do autor

Na página de visualização do autor pode-se observar uma lista de tópicos ordenada pela distribuição de frequência descrecente para o autor selecionado, uma lista de citações de sua autoria com o título, o periódico e o ano em que foi publicada. Também é apresentada uma lista dos 10 autores mais similares ao autor selecionado. Esta similaridade é baseada na distribuição de frequência de tópicos de cada autor e calculada por similaridade de cossenos entre o autor selecionado e os demais autores. Para cada artigo apresentado há um hiperlink para a página de visualização do artigo. Para cada tópico apresentado há um hiperlink para a página de visualização do tópico assim como para cada autor apresentado há um hiperlink para a página de visualização do autor (Figura 3).

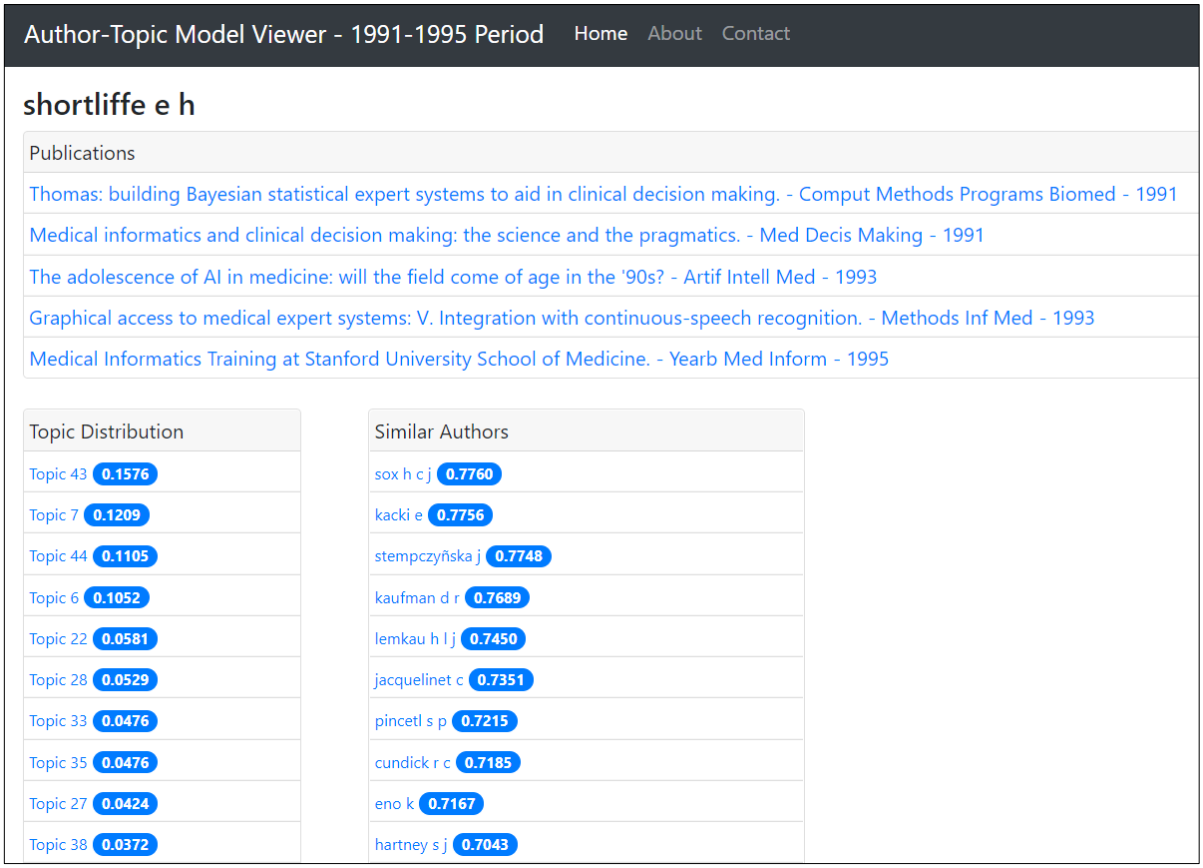


Figura 3 – Exemplo de recorte da página de visualização do autor

Página de visualização do artigo

Nesta página são apresentados o título, o periódico, o ano de publicação e o resumo do artigo. Também são apresentados os autores do artigo, em que para cada autor, é apresentada uma lista com os cinco tópicos com maior distribuição de frequência. Para cada autor apresentado há um hiperlink para a página de visualização do autor assim como para cada tópico apresentado há um hiperlink para a página de visualização do tópico (Figura 4).

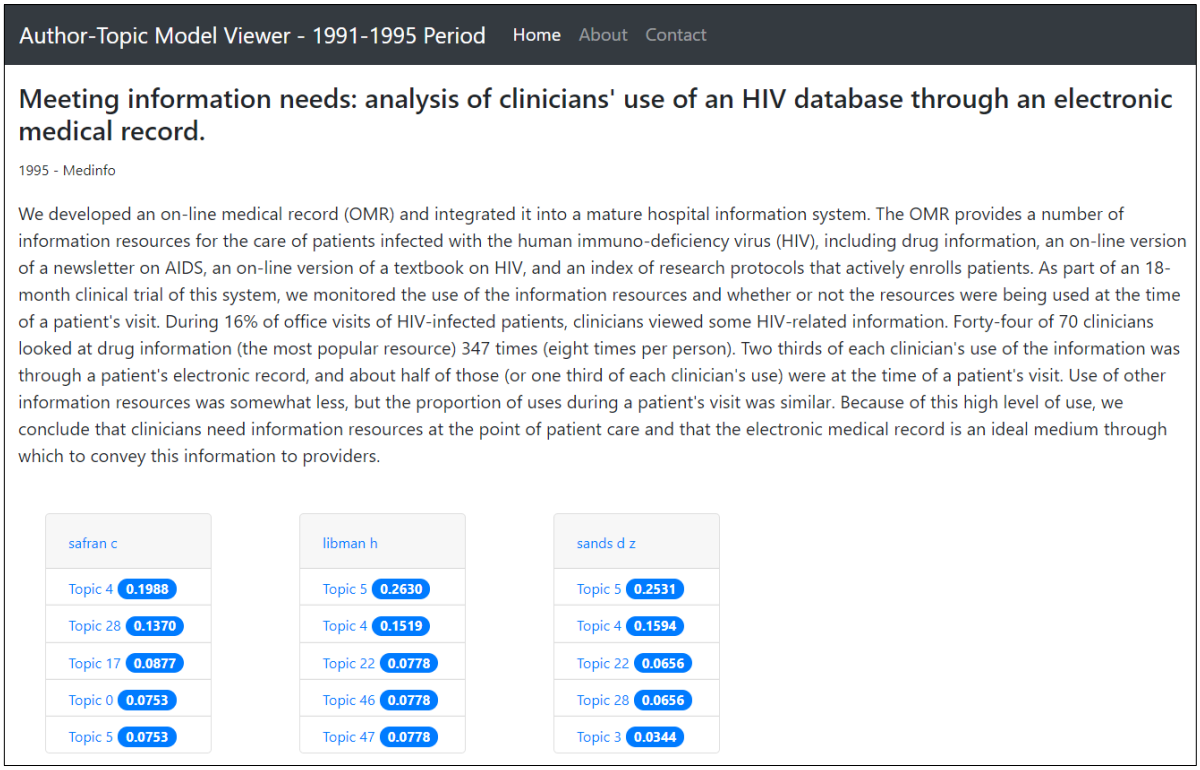


Figura 4 – Exemplo de página de visualização do artigo.

Anexo 5 – Resultados da rotulação de tópicos

Tópicos do período de 1991 a 1995			
Tópico	**Rótulo**	**Tópico**	**Rótulo**
0	representação do conhecimento	25	qualidade de imagens
1	NI	26	Sistema de apoio a decisão
2	processamento de imagens e sinais	27	programa de computador
3	análise estatistica	28	sistema de informação hospitalar
4	registro eletrônico do paciente	29	NI
5	registro eletrônico do paciente	30	padrões em IS
6	NI	31	PACS
7	yomada de decisão médica	32	análise de características de imagens
8	NI	33	gestão da informação em saúde
9	sistemas de gestãoo em saúde	34	pressão sanguínea
10	compressão de imagens	35	NI
11	análise estatística	36	NI
12	bioinformática	37	NI
13	testes diagnósticos	38	desenvolvimento de interface de usuário
14	engenharia biomédica	39	NI
15	NI	40	revisão sistemática
16	testes sanguíneos	41	NI
17	NI	42	NI
18	NI	43	educação em saúde
19	NI	44	representação do conhecimento
20	simulação matematica computador-izada	45	NI
21	hestão de sistemas de informação	46	NI
22	NI	47	processamento de dados
23	algoritmos baseados em filtro de ima-gens	48	análise de sinais
24	rede neural	49	sistemas de apoio a decisão

NI: não identificado

Tópicos do período de 1996 a 2000			
Tópico	**Rótulo**	**Tópico**	**Rótulo**
0	NI	25	NI
1	NI	26	NI
2	NI	27	NI
3	e-saúde	28	PACS
4	desenvolvimento de software	29	análise estatística
5	classificadores de padrões	30	sistemas de monitoramento
6	análise estatística	31	NI
7	detecção de imagens	32	NI
8	NI	33	telemedicina
9	survey	34	pressão sanguínea
10	bioinformática	35	NI
11	Processamento de imagens e sinais	36	NI
12	sistema de apoio a decisão clínica	37	Processamento de imagens e sinais
13	simulação de modelos baseada em computador	38	Educação em informática em saúde
14	qualidade de imagens	39	analise de custo benefício
15	Informação de saúde na web	40	engenharia biomédica
16	NI	41	sistema de apoio a decisão clínica
17	realidade virtual	42	ensaio clínico (clinical trial)
18	sistema de informação hospitalar	43	rede neural
19	educação médica	44	registro eletrônico de saúde
20	processamento de imagens e sinais	45	representação do conhecimento médico
21	sistema de informação hospitalar	46	realidade virtual
22	NI	47	análise de sinais
23	rede neural artificiais	48	Sistemas de gestão em saúde
24	NI	49	compressão de imagens

NI: não identificado

Tópicos do período de 2001 a 2005			
Tópico	**Rótulo**	**Tópico**	**Rótulo**
0	Processamento de imagens e sinais	25	sistemas de monitoramento
1	NI	26	coluna vertebral
2	análise de custo benefício	27	rede neural
3	realidade virtual	28	desenvolvimento de software
4	informação de saúde na web	29	processamento de imagens e sinais
5	desenvolvimento de software na web	30	análise estatística
6	registro eletrônico de saúde	31	simulação de modelos baseada em computador
7	rede neurl	32	sistemas de monitoramento
8	Survey	33	avaliação de software
9	bioinformática	34	NI
10	e-saúde	35	NI
11	Processamento de imagens e sinais	36	telemedicina
12	aprendizado de máquina	37	sistema de informação hospitalar
13	NI	38	análise estatística
14	NI	39	NI
15	NI	40	NI
16	processamento de imagens e sinais	41	educação médica
17	NI	42	revisão sistematica
18	saúde pública	43	NI
19	sistema de informação hospitalar	44	representação do conhecimento
20	mineração de dados	45	Quimioinformática
21	processamento de linguagem natural	46	gestão da saúde
22	Processamento de imagens e sinais	47	NI
23	NI	48	sistema de apoio à decisão clínica
24	NI	49	PACS

NI: não identificado

Tópicos do período de 2006 a 2010			
Tópico	**Rótulo**	**Tópico**	**Rótulo**
0	análise estatística	25	desenvolvimento de software na web
1	análise estatística	26	processamento de imagens e sinais
2	NI	27	processamento de linguagem natural
3	sistema de informação hospitalar	28	processamento de imagens e sinais
4	revisão sistemática	29	coluna vertebral
5	reconhecimento de padrões	30	avaliação e testes de confiabilidade
6	registro eletrônico de saúde	31	realidade virtual
7	sistema de apoio a decisão clínica	32	telemedicina
8	arquitetura de software	33	saúde pública
9	processamento de imagens e sinais	34	NI
10	gestão de TI em organizações de saúde	35	educação médica
11	NI	36	NI
12	NI	37	Quimioinformática
13	rede neural	38	desenvolvimento de software na web
14	processamento de imagens e sinais	39	análise estatística
15	survey	40	NI
16	bioinformática	41	usabilidade de software
17	simulação de modelos baseada em computador	42	diagnóstico por imagem
18	PACS	43	aprendizado de máquina
19	NI	44	survey
20	bioinformática	45	visualização de dados
21	processamento de imagens e sinais	46	informações de sáude na web
22	processamento de imagens e sinais	47	pressão sanguínea
23	cirurgia assistida por robô	48	análise de custo benefício
24	ontologia	49	representação do conhecimento

NI: não identificado

Tópicos do período de 2011 a 2015			
Tópico	**Rótulo**	**Tópico**	**Rótulo**
0	processamento de imagens e sinais	25	Processamento de imagens e sinais
1	NI	26	bioinformática
2	sistemas de apoio a decisão clínica	27	rede neural
3	NI	28	processamento de linguagem natural
4	cirurgia assistida por robô	29	otimização de algoritmos
5	segmentação de imagens	30	aprendizado de máquina
6	aprendizado de máquina	31	educação em saúde
7	revisão sistematica	32	sistemas de apoio a decisão clínica
8	saúde pública	33	segurança da informação em saúde
9	ensaio clínico	34	survey
10	detecção de imagens	35	análise estatística
11	informação de saúde para o consumi-dor	36	NI
12	NI	37	ontologias
13	pesquisa participativa baseada na comunidade (CBPR)	38	NI
14	Processamento de imagens e sinais	39	visualização de dados
15	usabilidade de software	40	e-saúde
16	monitoramento de saúde via celular	41	NI
17	NI	42	NI
18	simulação de modelos baseada em computador	43	rede neural
19	desenvolvimento de software	44	pressão sanguinea
20	registro eletrônico de saúde	45	bioinformática
21	ensaios clínicos baseados na web	46	aprendizado de máquina
22	realidade virtual	47	telemedicina
23	sistemade informação hospitalar	48	otimização de algoritmos
24	bioinformática	49	segmentação de imagens

NI: não identificado