Cristiane Raquel Woszezenki

MODELO PARA DESCOBERTA DE CONHECIMENTO BASEADO EM ASSOCIAÇÃO SEMÂNTICA E TEMPORAL ENTRE ELEMENTOS TEXTUAIS

Tese submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutor em Engenharia e Gestão do Conhecimento.
Orientador: Prof. Dr. Alexandre Leopoldo Gonçalves
Coorientador: Prof. Dr. João Artur de Souza

Florianópolis, SC
2016

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Woszezenki, Cristiane Raquel
Modelo para Descoberta de Conhecimento Baseado em Relações Semânticas e Temporais Entre Elementos Textuais / Cristiane Raquel Woszezenki; orientador, Alexandre Leopoldo Gonçalves; coorientador, João Artur de Souza. - Florianópolis, SC, 2016.
125 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Inclui referências

1. Engenharia e Gestão do Conhecimento. 2. Ontologia. 3.Associações. 4. Tempo. 5. Semântica. I. Gonçalves, Alexandre Leopoldo. II. de Souza, João Artur. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Cristiane Raquel Woszezenki

MODELO DE DESCOBERTA DE CONHECIMENTO BASEADO EM ASSOCIAÇÃO SEMÂNTICA E TEMPORAL ENTRE ELEMENTOS TEXTUAIS

Esta tese foi julgada adequada para obtenção do Título de "Doutor", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 08 de abril de 2016.

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

Banca Examinadora:

_____	_____
Prof. Alexandre Leopoldo Gonçalves, Dr. EGC/UFSC **Orientador**	Prof. Roberto Tadeu Raittz, Dr. Universidade Federal do Paraná
_____	_____
Profª. Renata Vieira, Dra. Pontifícia Universidade Católica do Rio Grande do Sul	Profª. Carina F. Dorneles, Dra. INE / UFSC
_____	_____
Prof. José Leomar Todesco, Dr. EGC / UFSC	Prof. Roberto Carlos dos Santos Pacheco, Dr. EGC / UFSC

À minha amada família,
especialmente meu filho Leonardo
e meu esposo Paulo.

AGRADECIMENTOS

À Deus, por ter criado e iluminado os caminhos que me conduziram a esta qualificação.

À minha família que tanto amo, de forma especial ao meu esposo Paulo que sempre me apoiou e ao meu filho Leonardo que veio ao mundo durante o andamento do doutorado me trazendo muita luz e me motivando ainda mais.

Ao meu orientador Alexandre Leopoldo Gonçalves, por ter me aceito como sua orientanda e, principalmente, por todo o apoio, atenção, dedicação, disponibilidade e auxílio que me deu. Muito obrigada Alex! Quero que Deus te abençoe o tempo todo.

Aos amigos e colegas Daniel Fernando Anderle e Vanderlei de Freitas Júnior pela companhia que tornaram as viagens Araranguá-Florianópolis e Florianópolis-Araranguá mais curtas e bem humoradas. Agradeço em especial ao Júnior pela parceria na publicação de artigos.

Ao Instituto Federal de Santa Catarina (IFSC), que me concedeu afastamento parcial do trabalho como professora oportunizando uma qualificação com certa tranquilidade.

Ao Programa de Pós Graduação em Engenharia e Gestão do Conhecimento (PPGEGC), pela formação.

Por fim, a todos os amigos que sempre me incentivaram.

“Tenho a impressão de ter sido uma criança brincando à beira-mar, divertindo-me em descobrir uma pedrinha mais lisa ou uma concha mais bonita que as outras, enquanto o imenso oceano da verdade continua misterioso diante de meus olhos”. (Isaac Newton)

RESUMO

O aumento da complexidade nas atividades organizacionais, a vertiginosa expansão da Internet e os avanços da sociedade do conhecimento são alguns dos responsáveis pelo volume inédito de dados digitais. Essa crescente massa de dados apresenta grande potencial para a análise de padrões e descoberta de conhecimento. Nesse sentido, a análise dos relacionamentos presentes nesse imenso volume de informações pode proporcionar novos e, possivelmente, inesperados *insights*. A presente pesquisa constatou a escassez de trabalhos que consideram adequadamente a semântica e a temporalidade dos relacionamentos entre elementos textuais, características consideradas importantes para a descoberta de conhecimento. Assim, este trabalho propõe um modelo para descoberta de conhecimento que conta com uma ontologia de alto-nível para a representação de relacionamentos e com a técnica *Latent Semantic Indexing* (LSI) para determinar a força de associação entre termos que não se relacionam diretamente. A representação do conhecimento de domínio, bem como, a determinação da força associativa entre os termos são realizadas levando em conta o tempo em que os relacionamentos ocorrem. A avaliação do modelo foi realizada a partir de dois tipos de experimentos: um que trata da classificação de documentos e outro que trata da associação semântica e temporal entre termos. Os resultados demonstram que o modelo: i) possui potencial para ser aplicado em tarefas intensivas em conhecimento, como a classificação e ii) é capaz de apresentar curvas da força associativa entre dois termos ao longo do tempo, contribuindo para o levantamento de hipóteses e, consequentemente, para a descoberta de conhecimento.

Palavras-chave: Engenharia do Conhecimento. Ontologia. Associações. Tempo. Semântica.

ABSTRACT

The increased complexity in organizational activities, the rapid expansion of the Internet and advances in the knowledge society are some of those responsible for the unprecedented volume of digital data. This growing body of data has great potential for pattern analysis and knowledge discovery. In this sense, the analysis of relationships present in this immense volume of information can provide new and possibly unexpected insights. This research found shortages of studies that adequately consider the semantics and the temporality of relationships between textual elements considered important features for knowledge discovery. This work proposes a model of knowledge discovery comprising a high-level ontology for the representation of relationships and the LSI technique to determine the strength of association between terms that do not relate directly. The representation of domain knowledge and the determination of the associative strength between the terms are made taking into account the time in which the relationships occur. The evaluation of the model was made from two types of experiments: one that deals with the classification of documents and another concerning semantics and temporal association between terms. The results show that the model: i) has the potential to be used as a text classifier and ii) is capable of displaying curves of associative force between two terms over time, contributing to the raising of hypotheses and therefore to discover of knowledge.

Keywords: *Knowledge Engineering. Ontology. Associations. Time. Semantics.*

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABC	Modelo de Descoberta Baseada em Literatura proposto por Swanson
DAML	DARPA *Agent Markup Language*
DBL	Descoberta Baseada em Literatura
DSRM	*Design Science Research Metodology*
EC	Engenharia do Conhecimento
GC	Gestão do Conhecimento
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Texts*
LSI	*Latent Semantic Indexing*
MD	Mineração de Dados
MeSH	*Medical Subject Headings*
MT	Mineração de Textos
NDE	Núcleo Docente Estruturante
NLM	*National Library of Medicine*
NMF	*Non-Negative Matrix Factorization*
OIL	*Ontology Inference Layer*
OWL	*Ontology Web Language*
PLN	Processamento de Linguagem Natural
PPGEGC	Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento
PSI	*Predication-based Semantic Indexing*
RDF	*Resource Description Framework*
RI	*Random Indexing*
SBC	Sistema Baseado em Conhecimento

SRI	*Symmetric Random Indexing*
SVD	*Singular Value Decomposition*
SVM	*Support Vector Machine*
UMLS	*Unified Medical Language System*

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

O aumento expressivo na quantidade de informação produzida e publicada no mundo tem sido cada vez mais discutido nos cenários acadêmico e empresarial. Isso se deve principalmente aos avanços das tecnologias da informação que acabam impactando na sociedade do conhecimento que, por sua vez, acabam por aumentar a quantidade de informações produzidas e facilitar o acesso dessas informações aos indivíduos e organizações.

As publicações científicas compreendem uma parcela significativa do volume inédito de dados digitais produzidos atualmente. O MEDLINE, por exemplo, é uma base de dados bibliográfica e disponibiliza mais de 22 milhões[1] de citações e resumos de publicações científicas das áreas da medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular, entre outros.

Além de artigos acadêmico-científicos e da Web, há ainda vários outros tipos de informação textual em formato digital disponíveis dentro das organizações: (a) os diversos tipos de relatórios; (b) manuais sobre procedimentos, softwares, etc.; (c) descrições textuais fornecidas por clientes sobre reclamações, elogios, ou sugestões sobre o produtos e/ou serviços; (d) os registros (arquivos de log) do sistema de busca textual da instituição ou mesmo de motores de busca, como o Google®, podem conter informações úteis sobre os interesses e necessidades dos seus colaboradores; (e) e-mails; (f) mensagens instantâneas; (g) currículos; (h) e-books; entre outros.

Esse cenário lança desafios sobre o armazenamento, a recuperação e a transformação dessas informações em conhecimento. Por outro lado, a superação desses desafios pode significar vantagem competitiva para as organizações. Nesse contexto, a Gestão do Conhecimento (GC), uma coleção de processos que governa a criação, disseminação e utilização do conhecimento para atingir plenamente os objetivos da organização (DAVENPORT; PRUSAK, 1998) tem se mostrado cada vez mais promissora. A Engenharia do Conhecimento (EC) também figura nesse cenário por ser a disciplina responsável pela criação de métodos e ferramentas que possibilitam o desenvolvimento de Sistemas Baseados em Conhecimento (SBC) para apoiar os diversos processos (criação, organização, formalização, compartilhamento,

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

aplicação e refinamento) da GC (STUDER et al., 1998; SCHREIBER et al., 2002).

A descoberta de conhecimento é um dos propósitos do desenvolvimento de SBCs e busca identificar ou explicitar novo conhecimento em um domínio de aplicação (KURGAN; MUSILEK, 2006). A descoberta de conhecimento pode ser realizada em bases de dados ou textos. A descoberta de conhecimento em bancos de dados dados (KDD, do inglês, *Knowledge Discovery in Database*) é o processo de descoberta de padrões e tendências por meio da análise de grandes conjuntos de dados, tendo como uma das etapas a Mineração de Dados (MD), que consiste na execução prática de análise e de algoritmos específicos que produz uma relação particular de padrões a partir de dados (FAYYAD et al., 1996). A descoberta de conhecimento em textos (KDT, do inglês, *Knowledge Discovery in Texts*) é o processo de descoberta de padrões e tendências por meio da análise de bases de dados textuais, ou seja, conjunto de dados não estruturados (FELDMAN et al., 1998). Uma das suas etapas é a Mineração de Textos (MT), responsável pela aplicação de algoritmos com o propósito de identificar padrões (SULLIVAN, 2001).

Uma importante vertente da MT é a descoberta de conhecimento por meio da análise dos relacionamentos entre elementos textuais (conceitos, termos, palavras, etc.).

A extração de relacionamentos é a principal atividade da mineração de textos biomédicos (WOSZEZENKI; GONÇALVES, 2013), que tem por objetivo detectar a ocorrência de relacionamentos entre genes, proteínas, genes e doenças, drogas e doenças, entre outros, buscando encontrar novos diagnósticos, prevenções e tratamentos. Os relacionamentos extraídos de textos biomédicos podem ser explícitos (diretos) ou implícitos (indiretos) (GANDRA et al., 2010). Relacionamentos explícitos são conhecidos e estão declarados no texto (por exemplo, “o gene G inibe a proteína P”). Já os relacionamentos implícitos são inferidos a partir da existência de um ou mais relacionamentos explícitos de forma a gerar hipóteses (que podem guiar experimentos em laboratórios) potenciais para novas descobertas (COHEN; HERSH, 2005). Este tipo de relacionamento é também chamado de associações (TSURUOKA et al., 2011).

As associações são objeto de estudo da Descoberta Baseada em Literatura (DBL), uma vertente da mineração de textos biomédicos. As descobertas se dão na forma de conexão implícita entre dois conceitos primários, como por exemplo, um medicamento para um gene que é a causa de uma doença (SWANSON, 1986; HRISTOVSKI et al., 2006).

Assim, a DBL pode ser utilizada para fazer a conexão entre conjuntos distintos de literatura, relacionando diferentes disciplinas ou especialidades (GANIZ et al., 2005). Suponha que uma comunidade científica sabe que B é uma das características da doença C. Outro grupo de pesquisadores sabe que a substância A afeta B. A descoberta, nesse caso, é o levantamento da associação de A com C por meio de B (WEEBER et al., 2001).

A DBL pode ser aplicada também em domínios diferentes do biomédico. Ittipanuvat et al. (2012), por exemplo, fazem uso da DBL para levantar conexões entre a tecnologia e questões sociais.

A descoberta de conexões entre domínios previamente desconectados também é foco de estudo da Bissociação. A Bissociação pode ser definida como conceitos (ou conjuntos de conceitos) que conectam dois domínios ainda não conectados, ou com conexão fraca/esparsa (BERTHOLD, 2012).

Existem outros cenários de usos potenciais das associações, como segurança nacional (SHETH et al., 2005), análise de redes sociais (ALEMAN-MEZA et al., 2006; OH; YEOM, 2012) e análise de histórico de falhas para evitar futuras falhas (GUO; KRAINES, 2010), detecção de ataques na web (KHAIRKAR et al., 2013), análise política (MOSCHOPOULOS et al., 2013), construção de mapas conceituais (KIM, 2013).

É importante ressaltar que, sob a perspectiva da Engenharia do Conhecimento, a associação pode ser vista como uma tarefa intensiva em conhecimento (SCHREIBER et al., 2002), uma vez que busca descobrir conhecimento implícito a partir de relacionamentos diretos.

1.2 IDENTIFICAÇÃO DO PROBLEMA

Conforme abordado anteriormente, as associações são inferidas a partir de um ou mais relacionamentos explícitos. Dessa forma, o sucesso na extração dos relacionamentos diretos tem impacto na geração de associações relevantes.

Durante muito tempo, o método mais utilizado para a extração de relacionamentos foi a coocorrência (SMALHEISER; SWANSON, 1998; HRISTOVSKI et al., 2001; EIJK et al., 2004; WU et al., 2004; ZHU et al., 2006; WREN, 2006; CHEN et al., 2008; TSURUOKA et al., 2011), o qual parte do pressuposto de que é possível identificar estatisticamente um possível relacionamento entre palavras, analisando suas frequências individuais e conjuntas. Contudo, esse tipo de abordagem não fornece explicitamente informações semânticas sobre a

natureza das relações entre os conceitos além de produzir um grande número de falsas relações (HRISTOVSKI et al., 2006; PREISS et al., 2012). Pesquisadores da área biomédica, por exemplo, não estão interessados apenas em relacionamentos baseados em coocorrência, mas também em entender os mecanismos de interação e causalidade desses relacionamentos (CAMERON et al., 2013). Assim, os pesquisadores passaram a incorporar a descrição semântica clara aos métodos e técnicas de forma a melhorar a qualidade das relações extraídas (HU et al., 2006, 2010; HRISTOVSKI et al., 2006; AHLERS et al., 2007; TSAI et al., 2007; BUNDSCHUS et al., 2008; KANG et al., 2011; LIU et al., 2012; KILICOGLU et al., 2012; GONG et al., 2013; CAMERON et al., 2013). A semântica de um relacionamento denota o significado desse relacionamento.

Há trabalhos que se baseiam nas relações semânticas presentes em ontologias ou terminologias biomédicas como o UMLS[2] (*Unified Medical Language System*) para extrair relacionamentos entre termos da literatura (RINDFLESCH et al., 2000; WEEBER et al., 2001; LEROY; CHEN, 2005; HRISTOVSKI et al., 2006; PATEL; CIMINO, 2009; HU et al., 2010; SUCHITRA; SUDHA, 2012; KILICOGLU et al., 2012). Hu et al. (2010), por exemplo, utilizam a coocorrência para a extração de termos, mas excluem termos para os quais não existe um relacionamento semântico no UMLS. O UMLS possui uma Rede Semântica a qual consiste de Tipos Semânticos, que categorizam os conceitos do UMLS, e um conjunto de Relações Semânticas existentes entre esses tipos. Os Tipos Semânticos são os nós da rede e as Relações Semânticas são links rotulados, como por exemplo, "*is_a*", "*associated_with*", "*part_of* ", "*affects*", "*treats*", "*prevents*", entre outros.

Além disso, diversos trabalhos buscam extrair predicações semânticas para conhecer a natureza das relações entre os termos (RINDFLESCH; FISZMAN, 2003; MCDONALD et al., 2004; AHLERS et al., 2007; HUNTER et al., 2008; HRISTOVSKI et al., 2010; WILKOWSKI et al., 2011; LIU et al., 2012; KILICOGLU et al., 2012; GONG et al., 2013; CAMERON et al., 2013). Predicações semânticas são relações binárias na forma Sujeito-Predicado-Objeto, onde o predicado (verbo) expressa o relacionamento entre o sujeito e o objeto (CAMERON et al., 2013). Contudo, Tsai et al. (2007) afirmam que a semântica dos relacionamentos vai além dos alvos principais da

[2] Metatesauro de conceitos médicos e relacionamentos semânticos entre eles, publicado e mantido pela NLM (*National Library of Medicine*), EUA. http://www.nlm.nih.gov/research/umls/

relação (sujeito e objeto) e o verbo que os conecta, ou seja, frases que descrevem localização, maneira, tempo, condições e amplitude são especialmente importantes no domínio biomédico. Os autores destacam que processos biológicos podem ser divididos em eventos temporais e espaciais, como por exemplo, a ativação de uma determinada proteína em uma célula específica ou a inibição de um gene por uma proteína em um determinado tempo. Para eles, ter informação compreensível sobre quando, onde e como esses eventos ocorrem é essencial para identificar as funções exatas de proteínas e a sequência de reações bioquímicas.

Embora existam diversos trabalhos focados na extração de relacionamentos semânticos entre termos, o nível de detalhamento semântico das relações ainda é baixo, o que dificulta a geração de associações relevantes.

Outro ponto faltante é o tratamento da dimensão temporal dos relacionamentos, sejam eles diretos ou indiretos. São poucos os trabalhos que se preocupam com os aspectos temporais dos relacionamentos (COHEN; SCHVANEVELDT, 2010; LOGLISCI, 2011; XU et al., 2013). Em geral, eles consideram os relacionamentos como sendo estáticos, ou seja, tentam determinar as conexões considerando a distribuição dos termos ou conceitos em um único ponto no tempo. Contudo, a grande maioria dos relacionamentos presentes em fontes textuais são intrinsecamente dinâmicos, ou seja, eles se alteram ao longo do tempo (LING; WELD, 2010; LOGLISCI, 2011). Por exemplo, o relacionamento entre duas pessoas pode mudar de orientador-orientando para colegas. A força do relacionamento também pode variar, podendo fortalecer-se ou enfraquecer-se. O relacionamento pode também ser finalizado, passando a não existir mais a partir de um determinado momento.

1.3 PERGUNTA DA PESQUISA

A partir das considerações iniciais e da identificação do problema de pesquisa, este trabalho busca responder à seguinte pergunta:

"Como as informações semânticas e temporais dos relacionamentos entre elementos textuais podem contribuir para a descoberta de conhecimento?"

1.4 PRESSUPOSTOS DA PESQUISA

Essa proposta de pesquisa possui os seguintes pressupostos:

- A riqueza de conteúdo presente nos dados digitais suscita o desenvolvimento de modelos de descoberta de conhecimento para apoiar os diversos processos da GC;
- A maior parte das informações disponíveis atualmente advém de fontes não estruturadas e métodos adequados devem ser desenvolvidos para tratar este tipo de informação;
- Os relacionamentos entre elementos textuais pode revelar conhecimento latente, isto é, conhecimento potencialmente existente, mas não facilmente evidenciado;
- A semântica presente nos relacionamentos favorece o entendimento aprimorado do domínio e pode tornar mais eficaz a descoberta de associações relevantes;
- Os aspectos temporais dos relacionamentos têm papel importante na compreensão da evolução dos domínios sob investigação;
- Diversos métodos, técnicas e ferramentas da EC podem ser integrados e combinados na proposição de um modelo de descoberta de conhecimento em fontes não estruturadas com ênfase na semântica e na evolução dos relacionamentos ao longo do tempo.

1.5 OBJETIVOS

A seguir são apresentados o objetivo geral e os objetivos específicos desta proposta de tese.

1.5.1 Objetivo Geral

Desenvolver um modelo de descoberta de conhecimento baseado em representação do conhecimento para extrair associações relevantes entre termos levando em conta os aspectos semântico e temporal dos relacionamentos.

1.5.2 Objetivos Específicos

Como objetivos específicos têm-se:

- Definir qual método de representação do conhecimento será utilizado na proposição do modelo;
- Definir qual técnica de associação será utilizada na proposição do modelo;

- Desenvolver um modelo para a extração de associações utilizando o método de representação do conhecimento e a técnica de associação identificados anteriormente;
- Demonstrar o funcionamento do modelo por meio de uma aplicação prática (nível de protótipo) em cenários de estudo;
- Realizar experimentos com o objetivo de avaliar o modelo.

1.6 JUSTIFICATIVA

Da grande e crescente quantidade de dados digitais em formato textual disponibilizada atualmente emerge a motivação inicial deste trabalho: o conhecimento latente existente a ser explorado neste vasto território. Em especial, a análise dos relacionamentos presentes em grandes bases textuais pode proporcionar novos e, possivelmente, inesperados *insights* (SHET et al., 2005), ou seja, a partir de dois ou mais relacionamentos diretos entre elementos textuais, podem ser derivadas novas conexões ou associações indicando conhecimento latente e, possivelmente, a descoberta de algo não evidenciado na literatura.

Outro importante aspecto que ressalta a relevância desse trabalho é o tratamento da semântica e temporalidade dos relacionamentos. A análise de relacionamentos e, consequentemente, a descoberta de novas conexões pode ser aprimorada pela exploração da semântica dos relacionamentos. Em outras palavras, o significado de um relacionamento possibilita um melhor entendimento sobre o domínio em investigação e, com isso, conexões mais relevantes podem ser levantadas.

A grande maioria dos domínios de conhecimento são dinâmicos, ou seja, mudam de acordo com o tempo. Dessa forma, o tratamento da temporalidade dos relacionamentos permite compreender a evolução desses relacionamentos de forma a melhor explicar um domínio dinâmico.

Assim, a proposta de um novo modelo de descoberta de conhecimento, que tenha por objetivo o levantamento de associações baseadas na semântica e na temporalidade dos relacionamentos entre elementos textuais justifica-se na medida em que busca preencher a lacuna identificada na literatura e contribuir com a área de pesquisa específica.

A seguir são apresentadas as características deste trabalho que o tornam original, relevante, viável e não trivial, bem como as suas principais contribuições.

1.6.1 Originalidade, relevância, viabilidade e não trivialidade

O caráter de **originalidade** deste trabalho é evidenciado pela revisão de literatura realizada, na qual não foram encontrados estudos que propusessem um modelo de descoberta de conhecimento capaz de representar detalhadamente a semântica e a temporalidade dos relacionamentos entre termos com aplicabilidade em qualquer domínio do conhecimento.

A **relevância** desta pesquisa está relacionada principalmente ao potencial de descoberta de conhecimento proporcionada por relacionamentos entre termos de documentos textuais em qualquer domínio. Os aspectos semântico e temporal dos relacionamentos são incorporados ao modelo proposto com o intuito de tornar mais eficaz o levantamento de associações relevantes.

A **viabilidade** fica igualmente demonstrada na medida em que as tecnologias, métodos e técnicas que integram o modelo proposto são consolidadas na comunidade científica da área de Engenharia do Conhecimento.

Por fim, as características reunidas no modelo, especialmente a representação das características semântica e temporal dos relacionamentos e a aplicação de decomposição matricial para o levantamento de associações demonstram sua **não trivialidade**.

1.6.2 Contribuições do modelo proposto

Destacam-se como principais contribuições desta pesquisa:

- Um modelo de descoberta do conhecimento baseado em associações semânticas e temporais;
- Um artefato genérico (ontologia de alto nível) para representação dos relacionamentos com ênfase nos aspectos semântico e temporal. Esse artefato pode ser estendido para representar conhecimento de diferentes domínios;
- Inserção da técnica LSI (*Latent Semantic Indexing*) ao modelo para maximizar o levantamento de associações relevantes, potencializando a descoberta do conhecimento;
- Demonstração do modelo por meio de aplicações em cenários de uso.

1.7 ADERÊNCIA AO OBJETO DE PESQUISA DO PROGRAMA

Este trabalho é uma pesquisa em nível de tese do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina – PPGEGC/UFSC.

O PPGEGC tem como objeto de pesquisa o conhecimento e os processos que o tornam fator gerador de valor na sociedade contemporânea.

Três epistemologias do conhecimento são apresentadas por Venzin et al. (1998):

Cognitivista: conhecimento é uma entidade (dados) fixa e representável, estocável em computadores, bases de dados, arquivos ou manuais e, portanto, conhecimento pode ser compartilhado em uma organização. São representantes desta escolha: Herbert Simon, Noam Chomsky e Marvin Minsky.

Conexionista: conhecimento está nas conexões de especialistas e é orientado à resolução de problemas. Conhecimento é interdependente da rede de componentes interconectados. Etienne Wenger, Bruce Kogut e Udo Zander estão inseridos nesta escola.

Autopoiética: conhecimento é resultado da transformação de informação feita pelo indivíduo, a partir de suas experiências e observações. Maturana e Varela e Nonaka e Takeuchi representam essa escola.

De acordo com Pacheco (2014), o PPGEGC considera que "conhecimento é conteúdo ou processo efetivado por agentes humanos ou artificiais em atividades de geração de valor científico, tecnológico, econômico, social ou cultural". Esta definição possui perfil interdisciplinar permitindo a convergência das três escolas acima citadas, ou seja, possui viés cognitivista, pois entende que o conhecimento pode ser conteúdo e estar presente em artefatos; conexionista, pois considera que rede e comunicação são essenciais na geração de valor; e possui também viés autopoiético, pois julga que a criação do conhecimento ocorre essencialmente por meio de humanos (PACHECO, 2014).

Este trabalho possui suas principais raízes na epistemologia cognitivista, contudo provê suporte para as abordagens conexionista e autopoiética.

A estruturação do PPGEGC em áreas de concentração é decorrente da atribuição de missões interrelacionadas aos processos de codificação/formalização (área de Engenharia); planejamento e gerência

(área de Gestão); e difusão, comunicação e compartilhamento (área de Mídia) do conhecimento (PACHECO et al., 2010). Dessa forma, o objeto de pesquisa e formação do PPGEGC é essencialmente interdisciplinar.

Esta tese está inserida na área da Engenharia do Conhecimento. No PPGEGC os objetivos da área da Engenharia do Conhecimento[3]

> "incluem a pesquisa e o desenvolvimento de técnicas e ferramentas para a formalização, codificação e gestão do conhecimento; de métodos de análise da estrutura e processos conduzidos por profissionais em atividades de conhecimento intensivo; e a pesquisa e desenvolvimento de sistemas de conhecimento."

Assim, a presente pesquisa está em consonância com os objetivos da Engenharia do Conhecimento, pois propõe um modelo de descoberta de conhecimento no qual o conhecimento de domínio é formalizado por meio de ontologias.

Também está em consonância com o caráter interdisciplinar do programa por propiciar a interação entre diferentes disciplinas, tais como Descoberta Baseada em Literatura, Engenharia do Conhecimento, Análise Temporal e Ciência da Computação em prol de uma proposição que objetiva colaborar com a Gestão do Conhecimento de organizações de diversos setores.

1.7.1 Teses e Dissertações do PPGEGC relacionadas

Com o intuito de comprovar a aderência da presente pesquisa ao objeto de estudo do PPGEGC, foi realizada uma pesquisa na base de teses e dissertações (http://btd.egc.ufsc.br) do Programa de forma a identificar dissertações e teses que possuam algum tipo de relação com este trabalho. Buscando pelo termo "ontologia", foram retornados 33 trabalhos que propõem, entre outras coisas: ontologias de domínio, ontologias que integram sistemas baseados em conhecimento, padrão de projetos de ontologias, método de construção de ontologias, entre outros. Assim, são listados a seguir os trabalhos que propõem ontologias que integram modelos de sistemas baseados em conhecimento por mais se assemelharem com a presente tese.

[3] http://www.egc.ufsc.br/pos-graduacao/programa/areas-de-concentracao/#ec

BUSS, Maico Oliveira. Modelo de Sistema de Conhecimento para Gestão de Listas de Espera para Cirurgias no Sistema Único de Saúde. Dissertação, 2015.

CECI, Flavio. Um Modelo Baseado em Casos e Ontologia para Apoio a Tarefa Intensiva em Conhecimento de Classificação com Foco na Análise de Sentimento. Tese, 2015.

NAZÁRIO, Débora Cabral. CUIDA – Um Modelo de Conhecimento de Qualidade de Contexto Aplicado aos Ambientes Ubíquos Internos em Domicílios Assistidos. Tese, 2015.

SILVA, Thales do Nascimento da. Um Modelo Baseado em Ontologia para Suporte a Tarefa Intensiva em Conhecimento de Recomendação. Dissertação, 2015.

BRAGLIA, Israel. Um Modelo Baseado em Ontologia e Extração de Informação como Suporte ao Design Instrucional na Geração de Mídias do Conhecimento. Tese, 2014.

KINCELER, Lucia Morais. Um Framework Baseado em Ontologia de Apoio à Gestão Estratégica da Inovação em Organizações de P&D+i. Tese, 2013.

SCHNEIDER, Viviane. Método de Modelagem do Contexto Estratégico para Sistemas baseados em Conhecimento. Dissertação, 2013.

ANDRADE, Rafael. Um modelo para recuperação e comunicação do conhecimento em documentos médicos. Tese, 2011.

BOVO, Alessandro Botelho. Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais. Tese, 2011.

HEINZLE, Roberto. Um Modelo de Engenharia do Conhecimento para Sistemas de Apoio a Decisão com Recursos para Raciocínio Abdutivo. Tese, 2011.

LOPES, Luiz Fernando. Um modelo de engenharia do conhecimento baseado em ontologia e cálculo probabilístico para o apoio ao diagnóstico. Tese, 2011.

RIBEIRO JÚNIOR, Divino Ignácio. Modelo de sistema baseado em conhecimento para apoiar processos de tomada de decisão em ciência e tecnologia. Tese, 2011.

STADNICK, Simone. Um Modelo de Conhecimento Para Uso de Balanço Hídrico Superficial no Apoio à Gestão de Recursos Hídricos. Dissertação, 2011.

BALANCIERI, Renato. Um método baseado em ontologias para explicitação de conhecimento derivado da análise de redes sociais de um domínio de aplicação. Tese, 2010.

CECI, Flávio. Um Modelo Semi-automático Para a Construção e Manutenção de Ontologias a partir de bases de documentos não estruturados. Dissertação, 2010.

MEDEIROS, Luciano Frontino de. Framework para Engenharia e Processamento de Ontologias utilizando Computação Quântica. Tese, 2010.

BEPPLER, Fabiano Duarte. Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico. Tese, 2008.

1.8 DELIMITAÇÕES E LIMITAÇÕES DO TRABALHO

Esta proposta de pesquisa pretende gerar um modelo de descoberta do conhecimento baseado na extração de associações levando em conta os aspectos semântico e temporal dos relacionamentos entre termos. Os termos dizem respeito a uma estrutura composta de uma ou mais palavras ou sigla que se refere a um determinado conceito em um domínio. Nesse sentido, deve-se ressaltar que não faz parte do escopo deste trabalho o desenvolvimento de técnicas e ferramentas para a extração de termos e de relacionamentos a partir dos textos. Assim, a produção de insumos fica sob responsabilidade dos usuários do modelo. Ainda, esta pesquisa não preocupa-se em determinar a estrutura das anotações dos documentos.

Embora este trabalho preocupa-se com a temporalidade dos relacionamentos, não é objeto de estudo desta pesquisa a lógica temporal. Esta pesquisa possui natureza interdisciplinar, não objetivando o aprofundamento de uma área específica.

Um protótipo foi desenvolvimento para avaliação do modelo proposto. Embora ferramentas e técnicas de visualização sejam importantes para a apresentação dos resultados obtidos, não faz parte do escopo desse trabalho o desenvolvimento de um produto com interface voltada ao usuário final.

1.9 ORGANIZAÇÃO DO TEXTO

O texto está organizado da seguinte forma: No capítulo 2 são apresentados os estudos teóricos sobre os temas que fazem parte do escopo desse trabalho: relacionamentos, semântica dos relacionamentos, associação, técnicas de associação, análise temporal das informações textuais, representação do conhecimento e trabalhos relacionados.

O Capítulo 3 apresenta o embasamento teórico acerca da metodologia dessa pesquisa, explicitando a distinção entre pesquisa científica e pesquisa tecnológica e apresentando a abordagem de pesquisa *Design Science Research Methodology*, empregada na condução desse trabalho.

O capítulo 4 aborda o modelo de descoberta de conhecimento proposto. O seu funcionamento é demonstrado por meio de sua aplicação em um cenário fictício do contexto acadêmico. Neste capítulo também é apresentada a ontologia de alto nível proposta para a representação de relacionamentos semânticos e temporais. A demonstração da aplicabilidade da ontologia é realizada por meio de um cenário do currículo Lattes.

O Capítulo 5 trata da avaliação do modelo perante dois tipos de experimentos. Na sequência são listadas as publicações realizadas durante o período de desenvolvimento desta tese e as referências bibliográficas utilizadas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma revisão da literatura dos temas que fazem parte do escopo desse trabalho: relacionamentos, semântica dos relacionamentos, associação, técnicas de associação, análise temporal das informações textuais e representação do conhecimento.

2.1 RELACIONAMENTOS

Neste trabalho, o termo *relacionamento* é utilizado para denotar uma relação direta e explícita entre termos.

A Biomedicina é o campo de estudo que mais tem desenvolvido trabalhos para a extração de relacionamentos entre termos. Nessa área, a detecção de relacionamentos entre genes, proteínas, genes e doenças, drogas e doenças, entre outros, proporciona a descoberta de novos diagnósticos, prevenções e tratamentos.

Os relacionamentos extraídos de textos biomédicos podem ser explícitos ou implícitos (GANDRA et al., 2010). Relacionamentos explícitos são conhecidos e estão declarados no texto (por exemplo, “o gene G inibe a proteína P”). Já os relacionamentos implícitos são inferidos a partir da existência de dois ou mais relacionamentos explícitos de forma a gerar hipóteses (que podem guiar experimentos em laboratórios) potenciais para novas descobertas (COHEN; HERSH, 2005). Este tipo de relacionamento é também chamado de associações (COHEN; HERSH, 2005; GUO; KRAINES, 2010; TSURUOKA et al., 2011). A associação é tarefa central a esta pesquisa e é abordada na seção 2.2.

Além da biomedicina, a extração de relacionamentos e associações tem outros casos de usos potenciais, como segurança nacional (SHETH et al., 2005), análise de redes sociais (ALEMAN-MEZA et al., 2006; OH; YEOM, 2012; MAGULURI et al., 2014), análise de histórico de falhas para evitar futuras falhas (GUO; KRAINES, 2010), detecção de ataques na web (KHAIRKAR et al., 2013), análise política (MOSCHOPOULOS et al., 2013), construção de mapas conceituais (KIM, 2013).

Durante muito tempo, o método mais utilizado para a extração de relacionamentos foi a coocorrência (SMALHEISER; SWANSON, 1998; HRISTOVSKI et al., 2001; EIJK et al., 2004; WU et al., 2004; ZHU et al., 2006; WREN, 2006; CHEN et al., 2008; TSURUOKA et al., 2011), o qual parte do pressuposto de que é possível identificar estatisticamente um possível relacionamento entre palavras, analisando suas frequências

individuais e conjuntas (GONÇALVES et al., 2006). Contudo, esse tipo de abordagem não fornece nenhuma informação semântica sobre a natureza das relações (HRISTOVSKI et al., 2006; PREISS et al., 2012) entre os conceitos, além de produzir um grande número de falsas relações (HRISTOVSKI et al., 2006). Mais recentemente tem-se incorporado a semântica aos métodos e técnicas de forma a melhorar a qualidade das relações extraídas.

2.1.1 A Semântica dos Relacionamentos

Semântica é uma disciplina linguística que tem por objeto a descrição das significações próprias às línguas e sua organização teórica (TAMBA-MECZ, 2006). Para Bruza (2006), o termo semântica deriva da intuição que as palavras observadas no contexto de uma dada palavra contribuem para o seu significado. A semântica está, portanto, associada ao significado, ao sentido e à interpretação de palavras, expressões ou símbolos.

A literatura que trata da semântica dos relacionamentos apresenta diversas classificações para os relacionamentos. Green et al. (2002) apresentam três metaclasses de relacionamentos: relacionamentos de equivalência, de hierarquia e de associação. Relacionamentos de equivalência representam os sinônimos ou quase sinônimos de um termo. Relacionamentos de hierarquia especificam termos genéricos e específicos. Hiponímia e meronímia são exemplos comuns de relacionamentos de hierarquia. O primeiro também é conhecido como relacionamento É-UM (o cão É-UM animal). O segundo expressa a relação PARTE-TODO (o motor é parte do carro). Relacionamentos de associação representam uma relação semântica não hierárquica, como por exemplo, a relação CAUSA-EFEITO (o acidente foi causado pelo motorista embriagado). Outras classificações são encontradas em (LANDIS et al., 1987; WINSTON et al., 1987; CHAFFIN et al., 1988).

Contudo, é importante destacar que não faz parte do escopo deste trabalho detalhar a classificação dos relacionamentos à luz da Ciência Linguística, pois o desenvolvimento de ferramentas para a extração automática de relacionamentos diretos e a identificação do seu tipo não constitui um objetivo dessa pesquisa. Nesse sentido, a semântica dos relacionamentos é concebida aqui como o significado dos relacionamentos, o qual confere um maior entendimento sobre o domínio investigado potencializando a descoberta de conhecimento.

No campo da Biomedicina, há trabalhos que se baseiam nas relações semânticas presentes no UMLS[4] (Unified Medical Language System) para extrair relacionamentos entre termos da literatura biomédica (RINDFLESCH et al., 2000; WEEBER et al., 2001; LEROY; CHEN, 2005; HRISTOVSKI et al., 2006; PATEL; CIMINO, 2009; HU et al., 2010; SUCHITRA; SUDHA, 2012; KILICOGLU et al., 2012). Hu et al. (2010), por exemplo, utilizam a coocorrência para a extração de termos, mas excluem termos para os quais não existe um relacionamento semântico no UMLS. O UMLS possui uma Rede Semântica a qual consiste de Tipos Semânticos, que categorizam os conceitos do UMLS, e um conjunto de Relações Semânticas existentes entre esses tipos. Os Tipos Semânticos são os nós da rede e as Relações Semânticas são links rotulados, como por exemplo, "*is_a*", "*associated_with*", "*part_of*", "*affects*", "*treats*", "*prevents*", entre outros.

Diversos trabalhos buscam extrair predicações semânticas para conhecer a natureza das relações entre os conceitos (RINDFLESCH; FISZMAN, 2003; MCDONALD et al., 2004; AHLERS et al., 2007; HUNTER et al., 2008; HRISTOVSKI et al., 2010; WILKOWSKI et al., 2011; LIU et al., 2012; KILICOGLU et al., 2012; GONG et al., 2013; CAMERON et al., 2013). Predicações semânticas são relações binárias na forma Sujeito-Predicado-Objeto, onde o predicado (verbo) expressa o relacionamento entre o sujeito e o objeto (CAMERON et al., 2013).

Tsai et al. (2007) afirmam que a semântica dos relacionamentos vai além dos alvos principais da relação (sujeito e objeto) e o verbo que os conecta, ou seja, frases que descrevem localização, maneira, tempo, condições e amplitude são especialmente importantes no domínio biomédico. Processos biológicos podem ser divididos em eventos temporais e espaciais, como por exemplo, a ativação de uma determinada proteína em uma célula específica ou a inibição de um gene por uma proteína em um determinado tempo. Ter informação compreensivel sobre quando, onde e como esses eventos ocorrem é essencial para identificar as funções exatas de proteínas e a sequência de reações bioquímicas.

Embora existam diversos trabalhos focados na extração de relacionamentos semânticos entre termos, percebe-se que o nível de detalhamento semântico das relações ainda é baixo.

[4] Metatesauro de conceitos médicos e relacionamentos semânticos entre eles. http://www.nlm.nih.gov/research/umls/

2.2 ASSOCIAÇÃO

Neste trabalho, enquanto o termo *relacionamento* é utilizado para denotar uma relação direta e explícita entre termos, conforme visto anteriormente, o termo *associação* refere-se a um relacionamento indireto e implícito levantado a partir de dois ou mais relacionamentos diretos. Conforme mencionado na seção anterior, a associação é central a esta pesquisa e é concebida como uma tarefa intensiva em conhecimento (SCHREIBER et al., 2002), visto que objetiva derivar novos conhecimentos a partir do domínio estudado.

Uma área de pesquisa muito difundida que faz uso da associação é a DBL, cujos trabalhos são em sua grande maioria aplicados em informações textuais das ciências biomédicas. Todavia, segundo (WEEBER, 2003), a DBL pode ser aplicada em textos de qualquer área de conhecimento.

A DBL foi introduzida por Don Swanson, na década de 1980, quando não existiam ferramentas de mineração automática de textos. Nesta época, as pesquisas manuais realizadas por Don Swanson na literatura científica resultaram na associação da "Síndrome de Raynaud" (uma condição que resulta em restrição intermitente do fluxo sanguíneo para os dedos, disparado pelo frio ou estímulos emocionais) com o "óleo de peixe" por meio da "alta viscosidade do sangue" (SWANSON, 1986). Inicialmente, Swanson revisou a literatura biomédica em busca de informações sobre a Síndrome de Raynaud. Seus estudos apontaram que pacientes com esta doença apresentavam alterações no sangue, como "alta viscosidade" e "elevada agregação de plaquetas". Na revisão da literatura sobre alta viscosidade do sangue, Swanson encontrou uma conexão entre esse termo e o termo "óleo de peixe", indicando que o óleo de peixe auxilia na diminuição da viscosidade do sangue e a agregação de plaquetas. Assim, ele encontrou um relacionamento indireto entre a doença "Síndrome de Raynaud" e os fatores "alta viscosidade do sangue" e "elevada agregação de plaquetas". Isso conduziu à geração da hipótese de que "óleo de peixe" pode ser útil para reduzir a "alta viscosidade do sangue" e a "elevada agregação de plaquetas" em seres humanos e então amenizar os sintomas da "Síndrome de Raynaud". Mais tarde, Swanson também encontrou um relacionamento entre os termos "Magnésio" e "Enxaqueca" (SWANSON, 1988).

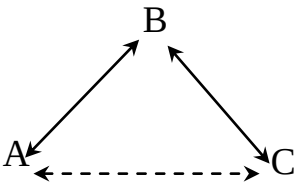
Posteriormente a essa importante descoberta, o surgimento de métodos e ferramentas computacionais de mineração de textos biomédicos alavancou a descoberta de conhecimento a partir da

literatura científica. Nesta seção são levantadas as principais técnicas da literatura utilizadas na tarefa de associação.

2.2.1 Modelo ABC

Desde que Swanson fez a sua primeira descoberta e introduziu o conceito da DBL, o modelo por ele utilizado ainda continua amplamente empregado pelos trabalhos da área. Tal modelo é chamado de ABC (WEEBER, 2003) e busca criar associações (relacionamentos indiretos) entre termos ou conceitos que não têm um relacionamento direto na literatura. Por exemplo, considerando a afirmação em um determinado artigo de que "A afeta B", e considerando a informação de que "B afeta C" apresentada por outro artigo, pode-se derivar a afirmação implícita de que "A afeta C" (Figura 1) (SMALHEISER, 2011). Ou seja, a premissa dessa abordagem é que existem duas disciplinas ou estruturas de conhecimento científico que não se comunicam diretamente. Contudo, parte do conhecimento de um domínio pode complementar o conhecimento de outro (WEEBER, 2003).

Figura 1: Modelo ABC de Swanson. Os relacionamentos AB e BC são conhecidos e apresentados explicitamente na literatura. O relacionamento AC é uma possível descoberta.

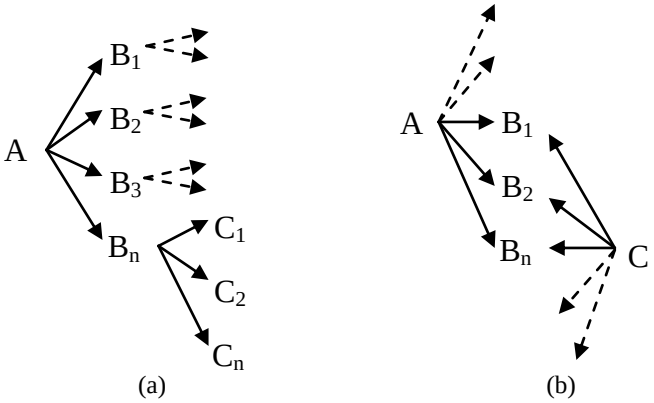


Fonte: (WEEBER, 2003)

Weeber et al. (2001) definiram duas abordagens de descoberta para o modelo ABC: fechada e aberta. A Figura 2 ilustra ambas as abordagens. A descoberta aberta é caracterizada como a geração de hipóteses como resultante do processo da DBL. Seu processo inicia com um conceito (um tópico ou um termo) conhecido, denotado como conceito inicial ou A. Um exemplo é uma doença. Em seguida, este conceito é utilizado na consulta a uma base de documentos, e todos os documentos contendo esse conceito são utilizados, compondo a literatura inicial. Termos importantes ou frases podem ser extraídos e

processados utilizando diferentes técnicas com o auxílio de um especialista humano. Cada termo ou frase da lista resultante é chamado de conceito intermediário ou B. Por exemplo, se A é uma doença (tal como “Doença de Raynaud”), então os conceitos Bs podem ser características ou sintomas da doença (como “agregação de plaquetas” e “viscosidade do sangue”). A literatura intermediária é processada e novos termos ou frases são extraídos novamente. Finalizando o processo, um conceito é selecionado por um especialista humano, denotado de conceito alvo ou C (como por exemplo, “óleo de peixe), gerando uma hipótese: C pode ser usado no tratamento da doença A em função de B.

Figura 2: Modelo de descoberta aberta (a) e modelo de descoberta fechada (b). As linhas contínuas indicam caminhos com potenciais de descoberta; as linhas tracejadas indicam caminhos sem sucesso.



Fonte: (WEEBER et al., 2001).

A descoberta fechada começa com A e C conhecidos. Podem ser uma associação observada, ou uma hipótese gerada previamente. A descoberta nesta situação concentra-se em encontrar novos Bs que podem explicar essa observação. É importante ressaltar que, em ambas as abordagens, tanto a aberta quanto a fechada, para que a descoberta seja efetivada, a hipótese formulada não pode ter sido explicitamente publicada.

2.2.2 Técnicas de Associação

Diversos trabalhos que buscam descobrir associações para gerar novas hipóteses fazem uso do modelo ABC aplicando abordagens baseadas em coocorrência para encontrar termos com relacionamentos diretos (AB e BC) e posteriormente identificar relacionamentos indiretos (AC) (SMALHEISER; SWANSON, 1998; WEEBER et al., 2001; SRINIVASAN, 2004; HU, 2005; HU et al., 2010). A maioria desses pesquisadores buscam melhorar os resultados dos seus trabalhos adicionando outros elementos as suas abordagens, como por exemplo, o refinamento dos relacionamentos e das associações com base nos tipos semânticos e relações semânticas presentes no UMLS. Contudo, esse tipo de refinamento é específico para o contexto médico, não podendo ser aplicado a outros contextos.

Alguns pesquisadores buscam explorar um maior número de termos intermediários para levantar associações entre A e C. Wilkowski et al. (2011), por exemplo, propuseram o modelo AnC, uma extensão do modelo ABC de Swanson, onde n = (B_1, B_2, ..., B_n). Para esses modelos que trabalham com mais de um termo intermediário, a teoria dos grafos apresenta-se como uma técnica promissora para descobrir novas associações. Um grafo é uma representação de conexões (arestas) entre objetos (nós). Grafos, também conhecidos como redes, são muito usados em análise de redes sociais e Web Semântica. A teoria dos grafos é um conjunto de funções e medidas pertinentes as propriedades dos grafos. Como exemplo de medida tem-se a centralidade, a qual mede o grau de conexão dos nós em um grafo. Um nó com mais conexões (relacionamentos) com outros nós tem um grau de centralidade maior. Freeman (1979) descreve o grau de centralidade como um indicador da atividade de comunicação em uma rede social. A importância da alta conectividade é também estudada em redes de interação genéticas (VOGELSTEIN et al., 2000). Existem trabalhos que propõem metodologias de DBL nas quais um grafo, cujas arestas rotuladas com predicações semânticas, possa ser navegado por um especialista de forma a encontrar caminhos que elucidem as relações, fornecendo novos pontos de vista sobre algum problema de pesquisa (BUNDSCHUS et al., 2008; GUO; KRAINES, 2010, 2011; WILKOWSKI et al., 2011; GOODWIN et al., 2012; GONG et al., 2013; CAMERON et al., 2013). Dupont et al. (2006) discutem várias abordagens de como percorrer um grafo extraindo caminhos. Anyanwu e Maduko (2005) exploram a noção de previsibilidade ao classificar caminhos de associações semânticas na Web Semântica. Nos resultados de seu trabalho,

caminhos mais longos revelam associações mais raras. No contexto de *Linked Data*, Vidal et al. (2010) apresentam uma solução baseada em grafos que modela a topologia da conexão dos dados e estima pesos que medem o quão importante e relevante são as associações entre dois termos.

Também no contexto da DBL, Cohen e Schvaneveldt (2010) fazem uso da SRI (*Symmetric Random Indexing*) para levantar associações entre termos ou conceitos sobre o tempo de forma a prever futuras conexões, o que torna esse trabalho fortemente relacionado a essa tese. Esse método é uma variante da RI (*Random Indexing*), o qual constrói um espaço semântico por meio de uma matriz incorporando uma descrição concisa das coocorrências entre termos e documentos. Sua abordagem é semelhante ao método LSI (*Latent Semantic Indexing*) (detalhado na próxima seção), um método de indexação e recuperação de informação que tem sido bem sucedido na tarefa de extração de relacionamentos indiretos entre termos e conceitos contidos em grandes coleções de documentos não estruturadas (IOANNOU et al., 2013). O objetivo da variante SRI é atingir a mesma redução de dimensionalidade da LSI utilizando métodos computacionais mais simples e menos onerosos, possibilitando assim a escalabilidade para corpus de textos maiores.

Regras de associação (AGRAWAL et al., 1996) também são bastante utilizadas por trabalhos que buscam gerar novas conexões na DBL. A extração de regras de associação é uma técnica de mineração de dados que gera regras do tipo "Se X Então Y" a partir de um banco de dados de transações, onde X e Y são conjuntos de itens que coocorrem em várias transações. No caso da DBL, o conjunto de documentos é considerado como um “banco de dados de transações”, onde cada documento será considerado uma “transação” e cada palavra é considerada um “item”. Hristovski et al. (2001) usam regras de associação entre pares de conceitos médicos como método para determinais quais conceitos estão relacionados a um conceito inicial. A regra na forma $X \rightarrow Y$ é avaliada, em geral, através das métricas de confiança e suporte, onde o suporte é a porcentagem de documento que contém X ou Y e a confiança indica, dentre os documentos que contém X, a porcentagem de documentos que também contém Y. Em outras palavras os autores usam o conceito de coocorrência como indicação de relação entre conceitos. Se X é uma doença, algumas possíveis relações podem ser: *tem_sintoma*, *é_causada_por*, *é_tratada_com*, etc. Contudo, eles não tratam da natureza da relação. Em outro trabalho, Hu et al. (2010) propõem regras de associação semânticas para reduzir o número

de regras. Eles usam a coocorrência pra criar as regras de associação, mas excluem aquelas regras cujos conceitos não possuem tipos semânticos ou relacionamentos no UMLS. Li et al. (2009) também buscam diminuir o número de regras de associação filtrando os conceitos de acordo com os descritores MeSH (*Medical Subject Headings*)[5]. Guo e Kraines (2010) extraem associações entre dois relacionamentos de tipos específicos que envolvam um conceito. Eles chamam isso de associação de relacionamento, que estabelece que "se A tem um relacionamento R1 com o conceito B, então é provável que A tenha um relacionamento R2 com o conceito C". A ênfase é na associação entre os relacionamentos (A → R1 → B) e (A → R2 → C) e não entre os conceitos. A justificativa para isso é que um relacionamento representa a forma pela qual um conceito modifica o outro semanticamente. Mais recentemente, Paul *et al*. (2014) introduziram um método de mineração de regras de associação que combina métricas de similaridade entre conceitos, formuladas usando a estrutura intrínseca de uma determinada ontologia, com medidas de interessabilidade tradicionais para calcular medidas de interessabilidade semânticas.

Na seção seguinte é apresentada pormenorizadamente a LSI (*Latent Semantic Indexing*), um método de indexação e recuperação de informação que tem sido bem sucedida na tarefa de extração de relacionamentos indiretos entre termos e conceitos contidos em grandes coleções de documentos não estruturadas (IOANNOU et al., 2013). Essa técnica foi escolhida para a composição do modelo de descoberta do conhecimento proposto na presente tese. A justificativa para a escolha dessa técnica também se encontra na seção seguinte.

2.2.3 *Latent Semantic Indexing*

Latent Semantic Indexing (LSI) é um método de indexação e recuperação de informações que busca extrair e representar o significado contextual de palavras por meio de análises estatísticas aplicadas a grandes corpus de textos (DEERWESTER et al., 1998; LANDAUER et al., 1998; BAEZA-YATES; RIBEIRO-NETO, 2001). Essa abordagem assume que há uma estrutura semântica oculta (latente) subjacente aos dados a qual possibilita melhorar a detecção de documentos relevantes para os termos presentes em consultas (DEERWESTER et al., 1998).

[5] O MeSH representa um tesauro de termos, publicado e mantido pela NLM, amplamente utilizado para o tratamento da informação em saúde e ferramentas de aplicação da informática e saúde. Disponível em: https://www.nlm.nih.gov/mesh/.

Uma matriz M é construída a partir de um corpus, na qual, cada linha representa um termo do conjunto de termos T e cada coluna representa um documento do conjunto de documentos D (Figura 3). Cada entrada da matriz é definida por uma função de peso, se $T_i \in D_j$, e zero, caso contrário. Essa matriz é chamada matriz termo-documento.

Figura 3: Exemplo de matriz termo-documento usando a função de peso *tf-id*. T1 é uma palavra que ocorre frequentemente em documentos, enquanto que T3, T4 e T6 são palavras mais raras e recebem um peso maior.

M	D_1	D_2	D_3	D_4	D_5	D_6	$\cdots$	D_n
T_1	0.00060	0.00012	0.00003	0.00003	0.00333	0.00048	$\cdots$	a_{1n}
T_2	0	0	0	0	0	0	$\cdots$	a_{2n}
T_3	0	2.98862	0	0	0	1.49431	$\cdots$	a_{3n}
T_4	0	0	0	13.32555	0	0	$\cdots$	a_{4n}
T_5	0	0	0	0	0	0	$\cdots$	a_{5n}
T_6	1.03442	1.03442	0	0	0	3.10326	$\cdots$	a_{6n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T_m	a_{m1}	a_{m2}	a_{m3}	a_{m4}	a_{m5}	a_{m6}	$\cdots$	a_{mn}

Fonte: (CHEN et al., 2013)

Em seguida, a LSI faz uso do modelo matemático *Singular Value Decomposition* (SVD), o qual constrói um espaço semântico onde termos e documentos fortemente associados são colocados um perto do outro. Segundo Deerwester et al. (1998), o SVD permite o arranjo do espaço de forma a refletir os padrões associativos mais relevantes nos dados, ignorando as influências menos importantes. De acordo com esses autores, nesse espaço, os termos que não apareceram em um documento podem acabar perto do documento, se isso for consistente com os principais padrões de associação nos dados. Eles afirmam ainda que dessa forma, a posição no espaço serve como um tipo de indexação semântica e a recuperação procede usando os termos de uma consulta para identificar um ponto no espaço e, por fim, os documentos em torno desse ponto são retornados ao usuário.

Para tanto, a SVD decompõe a matriz original termo-documento M em três matrizes:

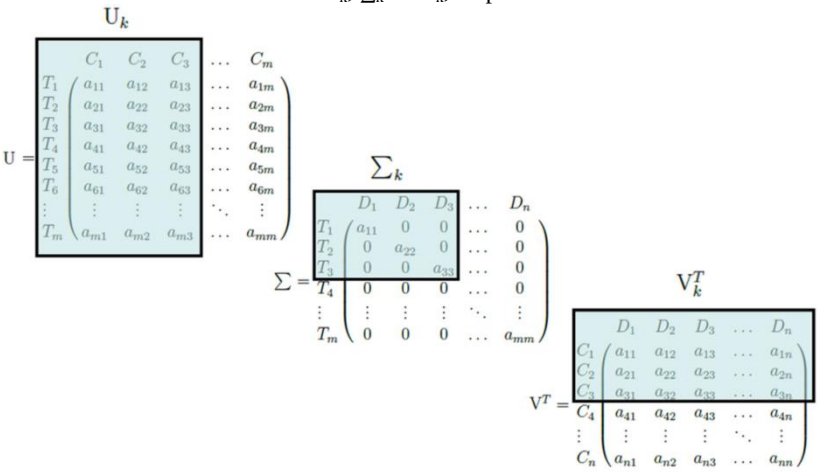
$$M = U \sum V^T$$

onde $\sum$ é uma matriz diagonal com as raízes quadradas dos autovalores de MM^T em ordem decrescente; U é uma matriz quadrada ortogonal de dimensões TxT com cada coluna representando o autovetor de MM^T, que corresponde a cada autovalor em $\sum$; e V^T é matriz transposta de uma matriz quadrada de dimensões DxD com cada coluna

representando o autovetor M^TM, que corresponde a cada autovalor em $\sum$.

A matriz U é a matriz termo-termo, com cada entrada representando uma relação de um termo com um conceito. De forma análoga, V^T é a matriz conceito-documento, com cada entrada representando a relação de um documento com um conceito. A LSI realiza a redução da dimensionalidade truncando cada matriz utilizando um parâmetro *k*. Os *k* valores mais singulares são obtidos de $\sum$, pois eles capturam o máximo da variância da matriz original e as primeiras *k* linhas e colunas são então obtidas das matrizes U e V^T, respectivamente. As matrizes resultantes U_k, $\sum_k$ e V^T_k, capturam a representação de M com dimensão reduzida (Figura 4).

Figura 4: Redução da dimensionalidade da matriz original M. As matrizes U, $\sum$ e V^T são truncadas e se tornam U_k, $\sum_k$ e V^T_k, respectivamente.



Fonte: (CHEN et al., 2013)

A decomposição gera uma matriz M_k aproximada à matriz original M, com espaço dimensional reduzido. Essa matriz M_k expressa a melhor representação da estrutura semântica de certo domínio. Ela deve ser pequena o suficiente para possibilitar rápida recuperação e grande o suficiente para representar adequadamente a estrutura semântica de um corpus (IOANNOU et al., 2013).

De acordo com Deerwester et al. (1998), podem ser realizadas três tipos de associações: termo-documento, termo-termo e documento-documento. Para a **associação termo-documento**, a equação aplicada é

a descrita acima ($M = U \sum V^T$). Para a **associação entre termos**, têm-se a seguinte equação:

$$MM^T = U \sum U^T$$

O produto escalar entre dois vetores-linha de M reflete a extensão na qual dois termos tem padrões de ocorrência similares no conjunto de documentos. A matriz MM^T é uma matriz quadrada simétrica contendo todos os produtos escalares termo a termo.

A **associação entre documentos** é similar, exceto que neste caso é o produto escalar entre dois vetores-coluna da matriz M que indica a extensão na qual dois documentos tem o mesmo padrão de termos. Assim, a matriz M^TM os produtos escalares documento a documento. A equação é a seguinte:

$$M^TM = V \sum V^T$$

Kontostathis et al. (2002) provam empiricamente que a LSI apresenta um caminho de conectividade para cada elemento não-zero na matriz termo-termo resultante. Eles explicam isso por meio de um exemplo, cujos dados foram obtidos de Deerwester et al. (1998). O Quadro 1 apresenta os títulos de cada documento.

Quadro 1: Títulos dos documentos.

c1:	Humam machine interface for Lab ABC computer applications
c2:	A survey of user opinion of computer system response time
c3:	The EPS user interface management system
c4:	System and human system engineering testing of EPS
c5:	Relation of user-perceived response time to error measurement
m1:	The generation of random, binary, unordered trees
m2:	The intersection graph of path in trees
m3:	Graph minors IV: Widths of trees and well-quase-ordering
m4:	Graph minors: A Survey

Fonte: (DEERWESTER et al., 1998)

O Quadro 2 mostra a matriz termo-documento para os títulos apresentados no Quadro 1. Cada entrada equivale ao número de ocorrências do termo no documento. O Quadro 3 apresenta a matriz termo-termo, cujas entradas equivalem à quantidade de vezes em que os termos coocorrem em um mesmo documento. No Quadro 4 tem-se a

matriz termo-termo recomposta após a aplicação do SVD, com a redução para 2 dimensões (k=2).

Quadro 2: Matriz termo-documento.

	c1	**c2**	**c3**	**c4**	**c5**	**m1**	**m2**	**m3**	**m4**
Human	1	0	0	1	0	0	0	0	0
Interface	1	0	1	0	0	0	0	0	0
Computer	1	1	0	0	0	0	0	0	0
User	0	1	1	0	1	0	0	0	0
System	0	1	1	2	0	0	0	0	0
Response	0	1	0	0	1	0	0	0	0
Time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
Survey	0	1	0	0	0	0	0	0	1
Trees	0	0	0	0	0	1	1	1	0
Graph	0	0	0	0	0	0	1	1	1
Minors	0	0	0	0	0	0	0	1	1

Fonte: (KONTOSTATHIS et al., 2002)

Quadro 3: Matriz termo-termo.

	Human	**Interface**	**Computer**	**user**	**System**	**Response**	**Time**	**EPS**	**Survey**	**trees**	**Graph**	**Minors**
Human	X	1	1	0	2	0	0	1	0	0	0	0
Interface	1	X	1	1	1	0	0	1	0	0	0	0
Computer	1	1	X	1	1	1	1	0	1	0	0	0
User	0	1	1	X	2	2	2	1	1	0	0	0
System	2	1	1	2	X	1	1	3	1	0	0	0
Response	0	0	1	2	1	X	2	0	1	0	0	0
Time	0	0	1	2	1	2	X	0	1	0	0	0
EPS	1	1	0	1	3	0	0	X	0	0	0	0
Survey	0	0	1	1	1	1	1	0	X	0	1	1
Trees	0	0	0	0	0	0	0	0	0	X	2	1
Graph	0	0	0	0	0	0	0	0	1	2	X	2
Minors	0	0	0	0	0	0	0	0	1	1	2	X

Fonte: (KONTOSTATHIS et al., 2002)

Quadro 4: Matriz termo-termo após redução de dimensionalidade.

	Human	**Interface**	**Computer**	**User**	**System**	**response**	**Time**	**EPS**	**Survey**	**Trees**	**graph**	**Minors**
Human	0.62	0.54	0.56	0.94	1.69	0.58	0.58	0.84	0.32	-0.32	-0.34	-0.25
Interface	0.54	0.48	0.52	0.87	1.50	0.55	0.55	0.73	0.35	-0.20	-0.19	-0.14
Computer	0.56	0.52	0.65	1.09	1.67	0.75	0.75	0.77	0.63	0.15	0.27	0.20
User	0.94	0.87	1.09	1.81	2.79	1.25	1.25	1.28	1.04	0.23	0.42	0.31
System	1.69	1.50	1.67	2.79	4.76	1.81	1.81	2.30	1.20	-0.47	-0.39	-0.28
Response	0.58	0.55	0.75	1.25	1.81	0.89	0.89	0.80	0.82	0.38	0.56	0.41
Time	0.58	0.55	0.75	1.25	1.81	0.89	0.89	0.80	0.82	0.38	0.56	0.41
EPS	0.84	0.73	0.77	1.28	2.30	0.80	0.80	1.13	0.46	-0.41	-0.43	-0.31
Survey	0.32	0.35	0.63	1.04	1.20	0.82	0.82	0.46	0.96	0.88	1.16	0.85
Trees	-0.32	-0.20	0.15	0.23	-0.47	0.31	0.38	-0.41	0.88	1.55	1.96	1.43
Graph	-0.34	-0.19	0.27	0.42	-0.39	0.56	0.56	-0.43	1.17	1.96	2.50	1.81
Minors	-0.25	-0.14	0.20	0.31	-0.28	0.41	0.41	-0.31	0.85	1.43	1.81	1.32

Fonte: (KONTOSTATHIS et al., 2002)

Kontostathis et al. (2002) consideram o valor da posição (*i, j*) como sendo o grau de associação entre o termo *i* e o termo *j* na coleção. Como pode ser visto no Quadro 4, os termos *user* e *human* agora tem o valor de 0.94, representando uma associação forte, onde antes o valor era zero (Quadro 3). Essa associação é resultado de uma relação transitiva: *user* tem relacionamento direto com *interface* e *interface* tem relacionamento direto com *human* (veja o Quadro 3). Essa relação transitiva é também conhecida como coocorrência de segunda ordem. Segundos os autores, isso demonstra o valor da abordagem LSI, uma vez que consultas com o termo *user* devem retornar documentos contendo o termo *human* no contexto dessa coleção .

Kontostathis et al. (2002) também observam o valor 0.15 no relacionamento entre *trees* e *computer*. Verificando os relacionamentos transitivos, obtém-se uma explicação sobre o porquê esses termos recebem um valor positivo (embora fraco) de associação. No Quadro 3 pode-se observar que *trees* coocorre com *graph*, *graph* coocorre com *survey* e *survey* coocorre com *computer*. Nesse caso, essa relação transitiva caracteriza uma coocorrência de terceira ordem.

2.2.3.1 Aplicações da LSI

De acordo com Ioannou et al. (2013), a LSI captura a estrutura subjacente de associações de dados e descobre padrões implícitos de forma eficiente. Ela tem sido validada experimentalmente (BERRY et al., 1994; DEERWESTER et al., 1998) e teoricamente provada como uma técnica eficaz para capturar a estrutura semântica inerente de um corpus (PAPADIMITRIOU et al., 1998).

A LSI tem sido usada de forma bem sucedida para formular novas hipóteses e gerar novas conexões a partir de conexões existentes (CHEN et al., 2013). Gordon e Dumais (1998), por exemplo, exploram a eficácia da LSI na DBL em dois processos: identificar relações "próximas" que são necessárias para iniciar o processo de descoberta, e descobrir relações mais distantes que podem gerar novas hipóteses de descoberta.

Kim et al. (2007) tentaram descobrir redes de interações entre genes utilizando LSI juntamente com NMF (*Non-Negative Matrix Factorization*), outro método de fatoração de matrizes. Os resultados mostraram que os métodos baseados em LSI e NMF superam os métodos de coocorrência.

Roy et al. (2011) demonstraram a capacidade da LSI em identificar conexões implícitas entre fatores de transcrição derivados de um conjunto de genes diferencialmente expressos.

Homayouni et al. (2005) exploraram a LSI para extrair relacionamentos entre genes a partir dos títulos e resumos do MEDLINE. Os resultados mostram que a LSI é um método robusto para elucidar relacionamentos explícitos e implícitos entre genes a partir da literatura biomédica. Essa característica torna essa técnica muito útil para a análise de novas associações em experimentos genômicos.

Abate et al. (2013) propõem a utilização da LSI para realizar a correlação entre genes e processos biológicos a partir das publicações científicas.

2.3 ANÁLISE TEMPORAL EM INFORMAÇÕES TEXTUAIS

Atualmente, a alta velocidade de produção de novos conteúdos torna a dimensão *tempo* uma propriedade intrínseca e relevante presente nas informações. Em muitos domínios encontram-se documentos textuais com alguma marcação de tempo (*timestamp*) associada. Pode-se citar como exemplo, notícias (dia da publicação), artigos científicos (ano

da publicação), mensagens de e-mails (dia do envio ou recebimento), etc. Padrões temporais interessantes podem ser extraídos dessas informações. Por exemplo, um assunto nos artigos de notícias geralmente tem uma estrutura temporal e evolucionária consistindo de temas (subtópicos) que caracterizam o começo, progresso, e impacto do evento. No caso de artigos científicos, o estudo de um tópico em algum período de tempo pode ter influenciado o estudo de outro tópico em época posterior (MEI; ZHAI, 2005). Dessa forma, análises temporais permitem ao usuário detectar associações importantes entre informações ao longo do tempo.

Assim, vários autores têm discutido a importância de se considerar a dimensão *tempo* na análise de informações textuais. Berrazega e Faiz (2013) afirmam que a *Web* se tornou a fonte de informação mais importante, cujo conteúdo descreve fatos e eventos dinâmicos de todos os tipos (políticos, culturais, econômicos, etc.) fazendo com que a extração de informações temporais seja um campo de pesquisa desafiador. Segundo Ha-Thuc et al. (2009), *padrões temporais* podem revelar informações úteis sobre o comportamento dos diversos tópicos nos conjuntos de dados. Khy et al. (2008) afirmam que trabalhos que tratam do processamento de documentos com ordem *temporal* são interessantes às áreas de recuperação e gestão da informação. He et al. (2010) consideram o entendimento sobre como tópicos na literatura científica *evoluem* um problema importante e interessante. Para Alonso et al. (2009), na medida em que a quantidade de informação aumenta rapidamente no mundo digital, o conceito do *tempo como uma dimensão* ao longo do qual a informação pode ser organizada e explorada torna-se cada vez mais importante.

A análise de tendências é um foco importante da mineração de informações textuais (LENT et al., 1997; MONTES-Y-GÓMEZ et al., 2001; FELDMAN; SANGER, 2006; CVIJIKJ; MICHAHELLES, 2011). Montes-y-Gómez et al. (2001) sugerem que a análise de tendências busca responder à questões como: Quais são as tendências gerais dos interesses da sociedade entre dois períodos? Há uma mudança significativa nos assuntos das notícias? Os assuntos são quase os mesmos nestes dois períodos ou são divergentes? Quais são os assuntos que estão emergindo ou desaparecendo? Algum assunto mantém o mesmo nível de ocorrência durante dois períodos? Entre outras. Dessa forma, a análise de tendências na mineração de textos depende da informação temporal dos documentos de uma coleção de forma que comparações possam ser realizadas entre um subconjunto de

documentos com relação a um período e outro subconjunto com relação a outro período (FELDMAN; SANGER, 2006).

A análise das relações entre elementos textuais no tempo também é um tema atacado por alguns autores. Cohen e Schvaneveldt (2010) consideram que a descoberta de relações implícitas (associações) em um conjunto de documentos demarcados pelo tempo pode ser vista como a previsão de futuras conexões explícitas. Eles afirmam que mudanças na força das associações de acordo com o tempo são importantes para prever conexões explícitas no futuro, uma vez que a força associativa entre dois conceitos a partir de campos diferentes tende a aumentar na medida em que novas conexões são descobertas entre outros conceitos nesses campos. Xu et al. (2013) propuseram um método para minerar e anotar relações semânticas temporais entre termos. Os autores afirmam que anotações semânticas temporais podem ajudar os usuários a descobrir e entender as relações semânticas desconhecidas ou emergentes entre termos.

Para Mengle e Goharian (2010), a descoberta de temas/categorias em evolução no tempo, bem como a evolução de seus relacionamentos, é um assunto de interesse em muitas aplicações. Subasic e Berendt (2008) ressaltam a necessidade de sistemas capazes de mostrar como tópicos emergem, evoluem e desaparecem (e talvez reaparecem) ao longo do tempo, e que técnicas de visualização são interessantes para mostrar os relacionamentos encontrados. Böttcher et al. (2008) afirmam que, no contexto da MT, é necessário o emprego de uma perspectiva com orientação temporal, colocando o entendimento das mudanças no centro da descoberta de conhecimento.

Diversos trabalhos tem se dedicado a extrair eventos, expressões temporais e relações temporais de textos clínicos (SAVOVA et al., 2009; RAGHAVAN et al., 2012; D'SOUZA; NG, 2013; SUN, W. et al., 2013; TANG et al., 2013). Segundo Sun et al. (2013), a dimensão temporal é essencial para a interpretação de narrativas clínicas. O autor cita que a progressão de doenças e eventos em um hospital geralmente são registrados cronologicamente e muitos eventos clinicamente relevantes são significativos apenas em um contexto temporal particular. Ele afirma ainda que a ordem na qual os sintomas se desenvolvem, o tempo dos tratamentos e a duração e frequência das medicações são importantes no contexto clínico.

Como percebe-se, a dinâmica temporal tem atraído muito interesse dos pesquisadores da mineração da literatura uma vez que ele tem um papel importante na compreensão da evolução dos domínios sob investigação. Contudo, constata-se que os estudos que buscam analisar a

evolução dos relacionamentos entre termos não apresentam um modelo de descoberta de conhecimento que seja aplicável a qualquer domínio e que dê ênfase na semântica dos relacionamentos.

2.4 REPRESENTAÇÃO DO CONHECIMENTO

A representação do conhecimento é um problema central da Inteligência Artificial e também uma importante atividade da Engenharia do Conhecimento. Brachman e Hill (1990) abordam a ideia da representação da seguinte forma: “como transmitir conhecimento a um robô ou outro sistema computacional de forma que, dada uma capacidade de raciocínio apropriada, aquele conhecimento possa ser usado pelo sistema para adaptar e explorar seu ambiente?”

McCalla e Cercone (1983) levantam algumas questões sobre a representação do conhecimento:

- Como estruturar o conhecimento explícito em uma base de conhecimento?
- Como codificar regras para manipular o conhecimento explícito de uma base e inferir conhecimento implícito?
- Como fazer e como controlar tais inferências?
- Como especificar formalmente a semântica de uma base de conhecimento?
- Como lidar com o conhecimento incompleto?
- Como extrair o conhecimento de um especialista para inicialmente alimentar a base de conhecimento?
- Como adquirir novos conhecimentos com o passar do tempo de forma que a base de conhecimento se mantenha atualizada?

De acordo com Vassev e Hinchey (2011), não existe uma classificação padrão de tipos de conhecimento, pois o domínio do problema é que determina quais tipos de conhecimento o projetista deve considerar e quais modelos ele pode derivar a partir daquele conhecimento. Os autores ainda sugerem que diferentes abordagens podem ser utilizadas e combinadas para representar diferentes tipos de conhecimento. Assim, neste trabalho são tratadas as principais abordagens para representação do conhecimento encontradas na literatura (MCCALLA; CERCONE, 1983; BRACHMAN; HILL, 1990; SOWA, 2000; BRACHMAN; LEVESQUE, 2004; VASSEV; HINCHEY, 2011): regras, frames, redes semânticas, lógica e ontologia.

Regras ou regras de produção consistem de duas partes: um conjunto antecedente de condições e um conjunto consequente de ações (BRACHMAN; LEVESQUE, 2004). Um exemplo dessa abordagem é o sistema especialista MYCIN (SHORTLIFFE, 1974), desenvolvido na década de 1970 para dar suporte ao diagnóstico e tratamento de infecção bacteriana . A maior vantagem dessa abordagem é a forma natural e simples de extrair, codificar conhecimento e entender o conhecimento codificado (MCCALLA; CERCONE, 1983; VASSEV; HINCHEY, 2011). Além disso, novas regras podem ser adicionadas e regras existentes podem ser excluídas independentemente de outras regras. Contudo, um sistema baseado em regras pode crescer muito incorporando milhares de regras e exigindo esforço extra e ferramentas para manter sua consistência (VASSEV; HINCHEY, 2011).

A noção de representação do conhecimento por meio de *frames* foi introduzida por Minsky (1975) e tem um papel importante na área. Um frame é uma estrutura de dados que inclui todo o conhecimento para representar um objeto, que pode ser um conceito ou uma situação. Ele também inclui diversos outros tipos de informações como por exemplo, informações descritivas, detalhes procedurais, entre outros. *Frames* são particularmente úteis quando usados para representar conhecimento de certos eventos ou conceitos estereotipados (MCCALLA; CERCONE, 1983). Segundo Luger (2004), embora os *frames* sejam uma ferramenta poderosa, muitos dos problemas de adquirir e organizar uma base de conhecimento complicada devem ser ainda resolvidos pela habilidade e intuição do programador.

Uma rede semântica é um grafo direcionado consistindo de nós – representando conceitos – conectados por arestas – que representam relações semânticas entre os conceitos (VASSEV; HINCHEY, 2011). De acordo com Mylopoulos (1980), a natureza das redes semânticas permite tratar diretamente questões de recuperação de informação, uma vez que associações podem ser usadas na busca por caminhos de acesso pela rede. Outra característica importante de esquemas de rede é o uso potencial de relações (*is_a*, *part_of*, etc.) para a organização de uma base de conhecimento. Além disso, a representação gráfica de bases de conhecimento em rede aumenta sua compreensibilidade. Mylopoulos (1980) menciona que a maior desvantagem de esquemas em rede é a falta de uma semântica formal e uma terminologia padrão.

Esquemas de representação do conhecimento baseados em lógica empregam as noções de constantes, variáveis, funções, predicados, conectores e quantificadores lógicos para representar fatos

(MYLOPOULOS, 1980). Uma base de conhecimento, de acordo com esse esquema, é uma coleção de fórmulas lógicas que fornecem uma descrição parcial do mundo. De acordo com Vassev e Hinchey (2011), a lógica é relevante, pois fornece uma semântica precisa e possibilita o raciocínio (inferência de novo conhecimento a partir do conhecimento existente), que por sua vez é relevante para a dedução. McCalla e Cercone (1983) ressaltam que a precisão formal e a interpretabilidade da lógica são úteis e fornecem expressividade que outros esquemas não o fazem.

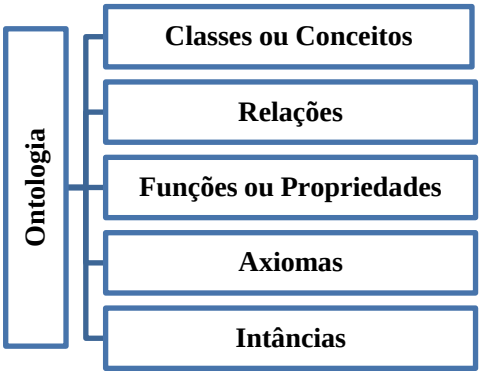
2.4.1 Ontologia

As ontologias são uma forma de organização e representação do conhecimento que vem recebendo especial atenção nos últimos anos devido à promessa de compartilhamento e entendimento comum de algum domínio de conhecimento que possa ser comunicado entre pessoas e computadores. Dessa forma, as ontologias têm sido desenvolvidas para propiciar o compartilhamento e reutilização de informações (GUARINO, 1995).

As ontologias tem sua estrutura baseada na descrição de conceitos e dos relacionamentos semânticos entre eles, gerando uma especificação formal e explícita de uma conceitualização compartilhada (STUDER et al., 1998). “Especificação explícita” diz respeito às definições de conceitos, instâncias, relações, restrições e axiomas; “formal” significa declarativamente definida, portanto compreensível para agentes e sistemas; “conceitualização” diz respeito a um modelo abstrato de parte do conhecimento; e “compartilhada” significa conhecimento consensual.

Para Gruber (1993) o conhecimento é formalizado na ontologia por cinco tipos de componentes: classes ou conceitos, relações, funções ou propriedades, axiomas e instâncias, conforme representado na Figura 5.

Figura 5: Estrutura de uma ontologia



Fonte: Baseado em (GRUBER, T. R., 1993)

Estes cinco componentes são comentados por Gómez-Pérez (1999), onde:

- *Classes*: são usadas em um sentido amplo. Um conjunto de classes e uma hierarquia entre estas classes formam uma taxonomia. Por exemplo, a classe “mãe” é uma subclasse da classe “mulher”;
- *Relações*: representam um tipo de interação entre as classes de um domínio. Um exemplo de relacionamento entre as classes “pessoa” e “casa” é o relacionamento “eh_proprietario”;
- *Funções*: é um caso especial de relacionamento em que um conjunto de elementos tem uma única relação com outro elemento. Um exemplo de função é “ser_pai”, onde a classe “homem” e a classe “mulher” estão relacionadas a uma classe “pessoa”;
- *Axiomas*: são regras que são sempre válidas. Um exemplo de axioma é afirmar que toda pessoa tem uma mãe. As regras possibilitam inferências sobre as classes;
- *Instâncias*: são especializações das classes, representam indivíduos específicos de uma determinada classe.

Guarino (1995) classifica as ontologias em 4 tipos:

- *Ontologias de alto-nível* – Descrevem conceitos gerais, tipicamente independentes de um problema particular ou domínio, como espaço, tempo, evento, etc. Ontologias desse

tipo são compartilhadas por grandes comunidades de usuários;

- *Ontologias de domínio* – Descrevem o vocabulário relacionado a um domínio genérico, como por exemplo, Direito, Engenharia, Biomedicina, entre outros. Podem especializar os conceitos das ontologias de alto nível.
- *Ontologias de tarefa* – Descrevem um vocabulário relacionado a uma tarefa ou atividade genérica, por meio da especialização de conceitos introduzidos nas ontologias de alto-nível.
- *Ontologias de aplicação* – São as ontologias mais específicas por serem utilizadas dentro das aplicações, especializando conceitos tanto das ontologias de domínio, quanto as de tarefas. Um exemplo é uma ontologia para uma aplicação que trabalhe com carros de luxo. Essa ontologia especializará conceito da ontologia de veículos (que é uma ontologia de domínio).

Para construir ontologias existem algumas linguagens formais específicas. Elas distinguem-se pelas facilidades, expressividades e propriedades computacionais que oferecem, sendo que as principais são: RDF (*Resource Description Framework*), RDF-Schema, OIL (*Ontology Inference Layer*), DAML (*DARPA Agent Markup Language*)+OIL e OWL (*Ontology Web Language*) (HEINZLE, 2011).

No processo de construção de ontologias são utilizados ambientes específicos que oferecem uma série de recursos e funcionalidades que auxiliam e facilitam o desenvolvimento do trabalho. Existem inúmeras ferramentas disponíveis que, embora tenham em comum o objetivo de oferecer facilidades para o desenvolvimento de uma ontologia, o fazem oferecendo diferentes recursos e funções. Duas das mais populares destas ferramentas são OntoEdit e Protégé (HEINZLE, 2011).

Heinzle (2011) também cita algumas propostas de metodologias de desenvolvimento de ontologias como a *Methontology*, a *On-to-Knowledge*, a *NeOn*, a *Uschold & King* e a *Grüninger & Fox*.

Segundo Gómez-Pérez (1999) e Rautenberg (2009), o desenvolvimento de ontologias é estudado na disciplina de Engenharia de Ontologias e baseado nas pesquisas de Gruber (1993) listam algumas recomendações:

- *Clareza*: uma ontologia deve claramente retratar e comunicar o significado dos elementos de um discurso, por meio de definições objetivas e bem documentadas;
- *Coerência*: em uma ontologia, quando existe uma lógica incutida, os axiomas devem ser consistentes, contribuindo para que as inferências geradas na utilização da ontologia estejam de acordo com o que se entende do domínio representado;
- *Extensibilidade*: as unidades de conhecimento de uma ontologia devem ser projetadas para que estas possam ser atualizadas e/ou reutilizadas. Em outras palavras, a extensibilidade diz respeito à incorporação de novos elementos, sem que os antigos necessitem ser revistos;
- *Limiar de codificação mínimo*: a conceitualização da ontologia deve ser especificada no nível de conhecimento, sem depender de uma linguagem específica. A linguagem específica de um domínio deve ficar no nível de instâncias da ontologia;
- *Compromisso ontológico mínimo*: uma ontologia deve definir apenas os termos extremamente suficientes para que as informações possam ser compartilhadas. Caso exista a necessidade de definições específicas para uma ontologia, reportando-se ao quesito reutilização de ontologias, uma ontologia pode ser instanciada e especializada para melhor descrever um domínio.

Gómez-Pérez (1999) afirma que o uso das ontologias permite a definição de um domínio no qual será possível trabalhar em determinada área específica, possibilitando a melhora no processo de extração de informação e o intercâmbio do conhecimento.

Guarino (1998) cita, entre outros benefícios das ontologias, a reutilização de vocabulários e a abstração de alto nível. A reutilização de vocabulários possibilita, em diferentes circunstâncias, a reutilização e o compartilhamento de conhecimento de domínios específicos através do uso de vocabulários comuns em todo o âmbito de determinada plataforma de *software*. A abstração de alto nível permite que os desenvolvedores se concentrem na estrutura a ser implementada sem a preocupação com o aprofundamento e excesso de detalhes de implementação.

Para Lopes (2011) o uso de uma ontologia permite a definição de um domínio no qual será possível trabalhar em determinada área específica, possibilitando a melhora no processo de extração de informação e o intercâmbio do conhecimento.

2.5 TRABALHOS RELACIONADOS

Do vasto campo da extração de associações, foram selecionados alguns trabalhos que buscam extrair associações semânticas, destacando-os como relacionados a esta tese. O Quadro 5 sintetiza estes trabalhos, caracterizando-os de acordo com: tempo (trata ou não da dimensão temporal das associações); representação do conhecimento (qual o tipo de representação do conhecimento é usada) e; abordagem (descrição breve sobre abordagem empregada no processo de descoberta).

Quadro 5: Trabalhos relacionados.

Trabalho	**Tempo**	**Repres. Conhec.**	**Abordagem**
Weissenborn et al. (2015)	Não	Grafos	Dados estruturados e não-estruturados são representados por um grafo, no qual, padrões de caminhos indicam associações biomédicas.
Cameron et al. (2015)	Não	Grafos e Tesauro MeSH	Dado um par de conceitos, o método proposto gera automaticamente uma lista ranqueada de subgrafos que fornecem informações e associações desconhecidas entre tais conceitos.
Xu et al. (2013)	Sim	Grafos	Grafos e PLN (Processamento de Linguagem Natural). Detalhes na Seção 2.5.1
Gong et al. (2013)	Não	Redes Semânticas e Ontologias Biomédicas	Usa abordagem baseada em PLN, analisando os verbos. Extraem todas as relações diretas e montam uma rede semântica, a partir da qual se consegue visualizar relações indiretas.
Cohen et al. (2012)	Não	Não	PLN e PSI (*Predication-based Semantic Indexing*)

			Identificam padrões do tipo "*drug x INHIBITS substance y, substance y CAUSES disease z"*.
Bovo (2011)	Sim	Ontologia	Coocorrência e similaridade entre vetores de contextos de dois conceitos. Detalhes na Seção 2.5.2.
Yang et al. (2011)	Não	Ontologia	PLN e SVM (*Support Vector Machine*) 1) Extraem entidades; 2) Extraem relacionamentos baseado em verbos; 3) Analisam a polaridade e força dos relacionamentos; e 4) Criam uma interface de visualização dos relacionamentos.
Guo & Kraines (2011)	Não	Ontologia e Grafos Semânticos	Criam um descritor para cada documento na forma de um grafo semântico. Os nós consistem de instâncias de conceitos da ontologia que representam entidades do documento; as arestas são relacionamentos que o artigo descreve entre as entidades. A abordagem possui um método que identifica uma associação em um determinado grafo e um algoritmo que extrai associações de um conjunto de grafos.
Sharma et al. (2010)	Não	Não	PLN. Dada uma sentença, extrai o verbo que indica um relacionamento e as entidades participantes do relacionamento.
Cohen e Schvaneveldt (2010)	Sim	Não	Coocorrência e SRI. Detalhes na Seção 2.5.3.

Hu et al. (2010)	Não	Tesauro UMLS	Coocorrência e Regras de associação. Usam a coocorrência pra extrair os termos, mas excluem termos cujos relacionamentos semânticos não existam no UMLS. Assim, conseguem diminuir o número de regras de associação.
Ahlers et al. (2007)	Não	Não	PLN. Extraem relacionamentos diretos do tipo Substance X <inhibits> Substance Y; Substance Y <causes> Pathology Z; E criam associações do tipo: Substance X <may_disrupt> Pathology Z.
Hristovski et al. (2006)	Não	Tesauro UMLS	PLN. É baseado em um formalismo gramático que combina sintaxe e semântica. A identificação das relações semânticas é baseada na rede semântica do UMLS.

Dentre os trabalhos relacionados acima elencados, três trabalhos são destacados por apresentarem-se fortemente relacionados a esta tese, pois, além de tratar da semântica das relações, buscam levantar associações levando em consideração a dimensão temporal dos relacionamentos. Todos são descritos a seguir.

2.5.1 Xu et al. (2013)

Xu et al. (2013) propuseram um método para minerar e anotar relações semânticas temporais entre termos ou elementos textuais, chamados por eles de entidades. Eles preocupam-se em fornecer a semântica e a temporalidade das relações. Os autores afirmam que um par de entidades pode ter diferentes relações semânticas em diferentes

intervalos de tempo. Neste trabalho são mineradas e anotadas as relações diretas e indiretas (associações).

Para isso, os pesquisadores constroem as entidades de conexão, padrões léxicos-sintáticos, sentenças de contexto, grafo de contexto e comunidades de contexto, descritos a seguir.

Entidades de conexão: Dado um par de entidades, podem ser identificadas as entidades que as conectam, refletindo sua relação implícita.

Padrão léxico sintático: É uma sentença (sequência de palavras) contendo o par de entidades. Por exemplo, "*A adquire B*", "*B é adquirido por A*".

Sentença de contexto: É uma sentença (sequência de palavras) que indica o contexto semântico no qual está inserido o par de entidades.

Grafo de contexto: É a estrutura de dados que reflete a relação entre entidades.

Comunidade de contexto: É um subgrafo do grafo de contexto que reflete uma parte do contexto do par de entidades.

Par de sentença: É um par de sentenças de contexto que refletem a relação implícita entre duas entidades. Por exemplo, as sentenças "*Andrew Tomkins, chief scientist of search, Yahoo®, USA*" e "*Evgeniy Gabrilovich, senior research scientist and manager of the NLP & IR Group at Yahoo®*" indicam que "Andrew Tomkins" e "Evgeniy Gabrilovich" tem relação do tipo "colegas".

Essas estruturas são construídas para cada intervalo de tempo. Nesse trabalho, os dados (entidades, sentenças, relações semânticas e tempo) são obtidos a partir de pesquisas ao Google®.

As relações diretas são mineradas a partir das entidades de conexão, padrão léxico sintático e sentenças de contexto. As relações indiretas (associações) são extraídas a partir das entidades de conexão e pares de sentenças.

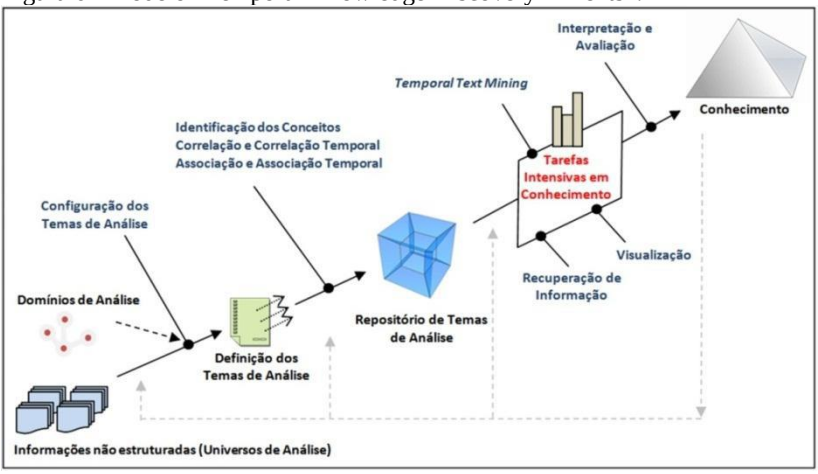
Os autores afirmam que anotações semânticas temporais podem ajudar os usuários a descobrir e entender as relações semânticas desconhecidas ou emergentes entre entidades.

2.5.2 Bovo (2011)

Bovo (2011) apresenta o modelo *Temporal Knowledge Discovery in Texts* (TKDT) com o objetivo de permitir a construção de sistemas de conhecimento que possibilitem aos usuários a execução de tarefas intensivas em conhecimento a partir da análise de informações

não estruturadas. Este modelo é iterativo e dividido por fases, conforme mostra a Figura 6.

Figura 6 - Modelo “Temporal Knowledge Discovery in Texts”.



Fonte: (BOVO, 2011)

Inicialmente realiza-se a configuração dos temas de análise**.** Um tema de análise consiste em um universo de análise (fontes de informação) e um domínio de análise (conhecimento de domínio). Em seguida localizam-se as ocorrências dos conceitos nos documentos textuais e marca-se um tempo (*timestamp*) a essa ocorrência. Após, criam-se as matrizes de correlação e associação.

Matriz de correlação: cada entrada dessa matriz representa a força do relacionamento direto entre dois conceitos (*i* e *j*), calculada a partir das suas coocorrências nos documentos textuais.

Matriz de associação: cada entrada dessa matriz representa a força do relacionamento indireto entre dois conceitos (*i* e *j*), calculada a partir da similaridade entre os vetores dos dois conceitos. O vetor de contexto de um conceito é formado pelos conceitos com os quais ele coocorre e seus respectivos pesos.

Às matrizes de correlação e associação *n* x *n* são adicionadas as seguintes dimensões:

Tempo (t): Tempo em que ocorre a correlação ou associação.

Tipo de relação (r): Essa dimensão pode ser utilizada para representar, por exemplo: i) As diferentes formas de coocorrências entre conceitos, tais como: coocorrência por documento, por janelas de

diferentes tamanhos, por sentença, etc.; ii) Relacionamentos que estejam definidos em uma ontologia do domínio de análise.

Tema de análise (a): Cada tema de análise é composto por um universo de análise e por um domínio de análise. O universo de análise corresponde às fontes de informação que serão utilizadas nas análises. Cada fonte de informação é formada por uma coleção de documentos textuais com algum atributo temporal como, por exemplo, a data de publicação. Já o domínio de análise refere-se ao conhecimento de domínio utilizado nas análises.

Assim, cada célula w_{ijklm} dessa matriz representa a força de correlação (ou associação) temporal entre dois conceitos (índices *i* e *j*), e um determinado tipo de relação (índice *k*), em um determinado tempo (índice *l*), para um determinado tema de análise (índice *m*). Essa matriz *n* x *n* x *r* x *t* x *a* representa o repositório dos temas de análise.

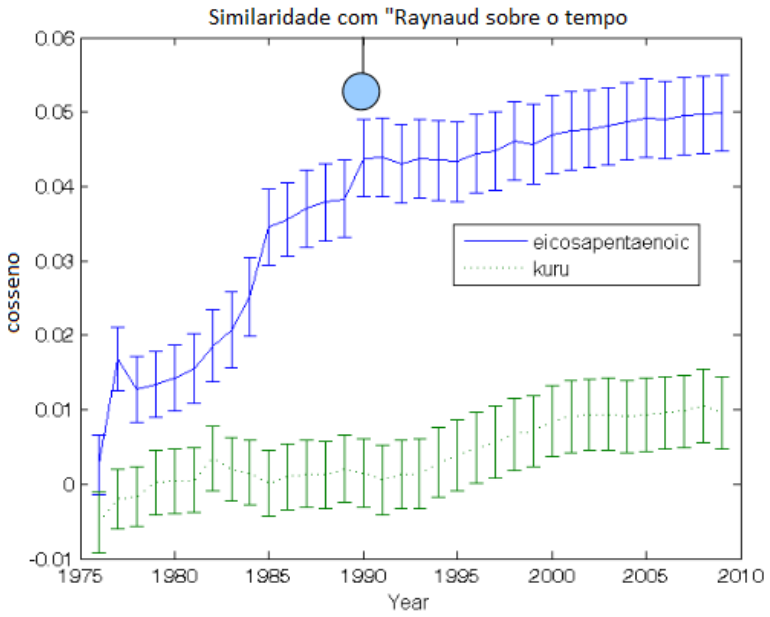
Bovo (2011) cita algumas tarefas que podem ser realizadas a partir do repositório de temas, como: DBL, Rastreamento de Tópicos, Análise de Relacionamentos Temporários, Detecção de Desvios, Extração de Regras de Associação Temporais e Visualização de Tendências.

2.5.3 Cohen & Schvaneveldt (2010)

Cohen e Schvaneveldt (2010) consideram que a descoberta de relações implícitas (associações) em um conjunto de documentos demarcados pelo tempo pode ser vista como a previsão de futuras conexões explícitas. Os autores propõem o uso da SRI, que analisa a coocorrência dos termos para determinar relacionamentos latentes (associações) entre os termos que não coocorrem. A SRI é semelhante à LSI, contudo, os autores mencionam o custo computacional dessa técnica e propõe a SRI com o intuito de diminuir a carga de processamento utilizando uma matriz reduzida, em vez de uma matriz esparsa.

Eles avaliam o modelo proposto por meio da reprodução das descobertas seminais de Swanson. Os termos são extraídos dos resumos do MEDLINE no período de 1975-2009. A força das associações é avaliada incrementalmente, ou seja, para cada ano são computadas as associações entre os termos. A Figura 7 ilustra a evolução da força associativa entre termos envolvendo uma das descobertas seminais de Swanson.

Figura 7: Associação entre os termos *raynaud-eicosapentaenoic* e *raynaud-kuru*.

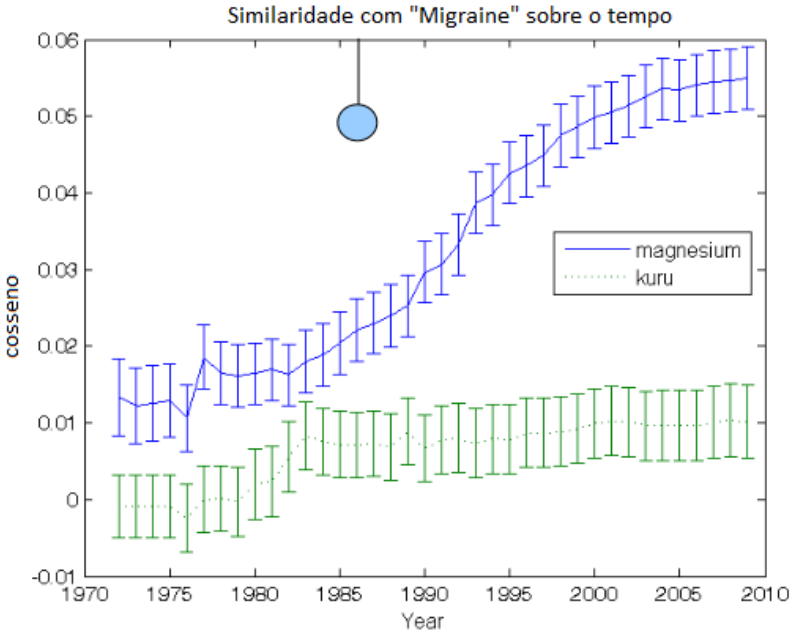


Fonte: Adaptado de (COHEN; SCHVANEVELDT, 2010).

A figura mostra um aumento na associação entre "*raynaud*" e "*eicosapentaenoic*". Esse comportamento precede a primeira vez em que esses termos coocorrem em um resumo do MEDLINE (círculo azul). Os pesquisadores observam que enquanto o primeira coocorrência observada ocorre aproximadamente cinco anos depois da descoberta de Swanson, o aumento na força associativa começa antes disso. O aumento da força associativa entre "*raynaud*" e "*eicosapentaenoic*" é maior do que entre "*raynaud*" e "*kuru*".

Outro comportamento semelhante a esse é observado entre os termos "*migraine*" e "*magnesium*", outra descoberta de Swanson conforme ilustra a Figura 8.

Figura 8:Associação entre os termos *migraine-magnesium* e *migraine-kuru*.



Fonte: Adaptado de (COHEN; SCHVANEVELDT, 2010).

Em ambos os casos, o aumento na força associativa precede a coocorrência direta. Assim, os autores concluem que a SRI é capaz de prever futuras coocorrências. Além disso, eles afirmam que essa técnica oferece escalabilidade para modelos em que as computações devem ser incrementais.

3 METODOLOGIA

A presente pesquisa, de acordo com Wazlawick (2010), tem a natureza original porque busca apresentar conhecimento novo a partir de observações e teorias construídas para explicá-lo. Em relação aos objetivos, ela é classificada como exploratória porque não pretende descrever os fatos, nem buscar suas causas e explicações, mas sim proporcionar maior familiaridade com o problema (GIL, 2008).

É também classificada, conforme Cupani (2011), Vargas (1985) e Bunge (1985) como pesquisa tecnológica, uma vez que se ocupa em projetar e desenvolver artefatos à luz do conhecimento científico. Esta modalidade de pesquisa é pautada pela tarefa que se propõe solucionar, tendo mais liberdade metodológica uma vez que a pesquisa tecnológica tem como produto, invariavelmente, o desenvolvimento de uma nova tecnologia.

A metodologia empregada para o desenvolvimento deste trabalho foi a *Design Science Research Methodology* (SIMON, 1996).

As seções a seguir apresentam o embasamento teórico acerca da metodologia dessa pesquisa, explicitando a distinção entre pesquisa científica e pesquisa tecnológica e apresentando a abordagem de pesquisa *Design Science Research Methodology*, empregada na condução desse trabalho.

3.1 PESQUISA CIENTÍFICA E PESQUISA TECNOLÓGICA

Cupani (2006) discute o conhecimento e a pesquisa tecnológica e científica. Para o autor, a tecnologia é o campo do conhecimento que se ocupa de projetar artefatos, planejar sua construção, operação, configuração, manutenção e acompanhamento, com base no conhecimento científico. Ele ainda observa que a tecnologia não é a mera aplicação do conhecimento científico, pois muitas das descobertas tecnológicas não surgiram a partir da aplicação da ciência.

Enquanto que o conhecimento científico propõe-se a desenvolver teorias de ampla aplicação, o conhecimento tecnológico é responsável pelo desenvolvimento de teorias de aplicação extremamente restritas, com vistas à solução de problemas pontuais e na maioria das vezes isolados, mais voltados à inovação tecnológica.

Cupani (2006) afirma que a tecnologia pode ser entendida a partir de quatro perspectivas básicas, a saber: a) como certo tipo de objeto (os artefatos); b) como uma classe específica de conhecimento (o saber tecnológico); c) como um conjunto de atividades para a produção e uso

de artefatos; e d) como uma manifestação de determinada vontade do ser humano em relação ao mundo.

Bunge (1985) afirma que a tecnologia pode ser vista como "o campo do conhecimento relativo ao projeto de artefatos e ao planejamento de sua realização, operação, ajuste, manutenção e monitoramento, à luz do conhecimento científico" (BUNGE, 1985, p. 231).

Enquanto que o conhecimento científico proporciona teorias mais abrangentes, o conhecimento tecnológico desenvolve teorias mais limitadas, que se propõem a atingir um problema específico, implicando sempre em invenção (CUPANI, 2006).

O conhecimento científico é constituído de teorias de amplo alcance e utiliza-se de idealizações, obrigando inclusive a sua adaptação para permitir sua aplicação. Já as teorias empregadas nas pesquisas tecnológicas são de aplicação limitada, uma vez que o conhecimento tecnológico é específico para uma determinada tarefa (CUPANI, 2006).

Enquanto que a pesquisa científica visa a descoberta de algo existente, a pesquisa tecnológica busca produzir algo novo (CUPANI, 2006). O produto da ciência (o conhecimento) é neutro (a teoria da evolução, por exemplo, não é boa nem má). Já os produtos da tecnologia podem ser benéficos, como a máquina de costura e a calculadora de bolso; ou maléficos, como um avião bombardeiro, a cadeira elétrica; ou ainda ambivalente, como o automóvel, a televisão e a aviação (CUPANI, 2011).

Freitas Junior *et al*. (2014) realizam uma distinção entre a pesquisa científica e tecnológica. O Quadro 6, apresentado pelos autores, cita e descreve as principais características destas duas formas de pesquisa.

Quadro 6: Pesquisa científica e pesquisa tecnológica.

Característica	**Pesquisa Científica**	**Pesquisa Tecnológica**
Definição	Conhecimento da natureza e exploração desse conhecimento (KNELLER, 1980).	"O estudo científico do artificial". "Tecnologia pode ser vista como o campo do conhecimento relativo ao projeto de artefatos e ao planejamento de sua realização, operação,

		ajuste, manutenção e monitoramento, a luz do conhecimento científico." (BUNGE, 1985).
Teorias	Amplo alcance e uso de idealizações, o que obriga a adaptar o conhecimento científico para possibilitar sua aplicação (CUPANI, 2006).	Aplicação limitada, pois o conhecimento tecnológico é específico pra uma determinada tarefa. Dois tipos: substantivas (conhecimento sobre a ação tecnológica) e operativas (conhecimento sobre as ações de que dependem o funcionamento dos artefatos) (CUPANI, 2006).
Resultado	Descobrimento de algo existente. O produto é neutro (nem bom nem mau) (CUPANI, 2006; CUPANI, 2011).	Criação de algo novo. O produto não é nem pode ser neutro. É, no mínimo, ambivalente (CUPANI, 2006; CUPANI, 2011).
Conhecimento	Descritivo (CUPANI, 2006).	Prescritivo. Específico. Peculiar. Conhecimento tácito, do saber-como. (CUPANI, 2006; CUPANI, 2011).
Desafios		Factibilidade, confiabilidade, eficiência dos inventos, relação custo-benefício (CUPANI, 2006).
Limitação	Ditada pela teoria.	Ditada pela tarefa

	Pode-se explorar livremente as possibilidades (CUPANI, 2006; CUPANI, 2011).	Imposta (CUPANI, 2006).
Origem dos dados	Científicos (CUPANI, 2006).	Experiência não científica (CUPANI, 2006). Dados relativos às exigências (técnicas, econômicas, culturais) que o artefato deve satisfazer (CUPANI, 2011).
Tipos de leis	Leis que governam os fenômenos naturais (CUPANI, 2006).	Regras de ação para dar origem aos fenômenos artificiais (CUPANI, 2006).
Pensamento	Abstrato e verbal (CUPANI, 2006)	Analógico e visual. (CUPANI, 2006)
Origem das variáveis	Não específico. (CUPANI, 2006).	Metas a alcançar (CUPANI, 2006).
Objetivos dos experimentos	Entender a realidade (CUPANI, 2006).	Conhecimento prático: "o artefato funcionará?", "haverá, acaso, fatores não previstos teoricamente que serão detectados experimentalmente?" (CUPANI, 2006). Controlar a realidade (CUPANI, 2006).
Explicações	Causais (CUPANI, 2006).	Funcionais (CUPANI, 2006).
Noção de conhecimento	Muda de acordo com as teorias (CUPANI, 2006).	Admitem apreciação de sua verdade ou falsidade, podendo-se afirmar que o artefato

		desempenha bem ou mal sua função. Superior em relação ao científico por sua certeza e eficácia (CUPANI, 2006).
Mudança de paradigma	Implica em muito exame e discussão (CUPANI, 2011).	Ocorre devido a anomalias funcionais ou presumíveis. A necessidade da mudança é percebida mais diretamente (CUPANI, 2011).
Revoluções	Inovadoras e eliminatórias (CUPANI, 2011).	Não implicam necessariamente em uma seleção radical, não supõem forçosamente uma nova comunidade e são compatíveis com a continuidade da tecnologia "normal" (CUPANI, 2011).

Fonte: FREITAS JUNIOR et al. (2014)

3.2 *DESIGN SCIENCE RESEARCH METHODOLOGY* (DSRM)

A Metodologia de Pesquisa da Ciência do *Design* (*Design Science Research Methodology*) tem sua origem na diferenciação entre os ambientes naturais e artificiais proposta por Simon (1996). Para o autor, a ciência natural se ocupa de descrever e ensinar como os fenômenos naturais funcionam e interagem com o mundo e as ciências do artificial se ocupam da concepção de artefatos que realizem objetivos.

Neste contexto, Simon (1996) argumenta pela necessidade de criar uma ciência que se dedique a propor como construir artefatos que possuam certas propriedades desejadas, ou seja, como projetá-los. Tal é o que ele chamou de *Design Science*, ou Ciência de Projeto. Segundo o autor, ao projeto cabe aspectos de “o que” e “como” as coisas devam

ser, e especialmente, a concepção de artefatos que tenham por propósito a realização de objetivos.

Segundo Lacerda et al. (2013), a *Design Science* é a base epistemológica e a *Design Science Research* é o método que operacionaliza a construção do conhecimento neste contexto. Para Çağdaş e Stubkjær (2011) a "*Design Science Research* se constitui em um processo rigoroso de projetar artefatos para resolver problemas, avaliar o que foi projetado ou o que está funcionando, e comunicar os resultados obtidos".

De acordo com March e Smith (1995), os artefatos podem ser:

- **Constructos**: Conceitos que formam o vocabulário de um domínio, definindo os termos usados para descrever e pensar sobre as tarefas.
- **Modelos**: Conjunto de proposições que expressam relacionamentos entre constructos. Modelos podem ser vistos como uma descrição ou uma representação de como as coisas são.
- **Métodos**: Conjunto de passos usados para executar determinada tarefa.
- **Instanciações**: É a concretização de um artefato em seu ambiente, demonstrando a viabilidade e a eficácia dos modelos e métodos. As instanciações informam como implementar ou utilizar determinado artefato e seus possíveis resultados.

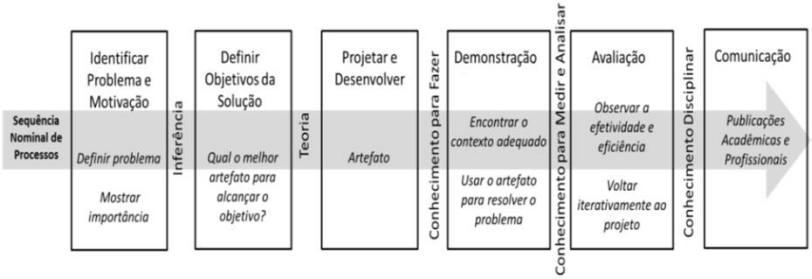
A DSRM, segundo Peffers et al. (2007), é desenvolvida a partir de seis etapas procedurais, que são sugeridas na ordem especificada na Figura 9, mas que podem ser executadas de acordo com a necessidade de projeto:

- **Identificação do problema e sua motivação**: esta etapa é dedicada à definição do problema de pesquisa específico, apresentando-se uma justificativa para a sua investigação. É importante que a definição deste problema seja empregada na construção de um artefato que pode efetivamente oferecer a solução para este problema. Tem-se como recursos necessários para esta etapa o estado da arte do problema e da relevância da solução apresentada.
- **Definição dos objetivos para a solução:** tendo-se como ponto de partida o conhecimento acerca do problema, bem como a noção do que é viável e factível, delineiam-se os

objetivos da solução a ser desenvolvida. Elencam-se como requisitos desta etapa novamente o estado da arte do problema e o conhecimento das possíveis soluções já previamente apresentadas.

- **Projetar e desenvolver**: etapa destinada à criação do artefato, determinando-se a sua funcionalidade desejada para o artefato, sua arquitetura e em seguida a criação do próprio artefato. Os recursos necessários para a terceira etapa compreendem o conhecimento da teoria que pode ser exercida em uma solução.
- **Demonstração**: momento de demonstração do uso do artefato resolvendo uma ou mais instâncias do problema por meio de um experimento ou simulação, estudo de caso, prova formal ou outra atividade apropriada. Os recursos relacionados para esta etapa incluem o conhecimento efetivo de como usar o artefato para resolver o problema.
- **Avaliação**: nesta etapa deve-se observar e mensurar como o artefato atende à solução do problema, comparando-se os objetivos propostos para a solução com os resultados advindos da utilização do artefato. Nesta etapa, pode-se definir pela recursividade da metodologia, isto é, o retorno às etapas Projetar e desenvolver ou Demonstração, de modo a aprimorar o artefato.
- **Comunicação**: momento de divulgação do problema e da relevância da propositura de uma solução para ele, além da apresentação do artefato desenvolvido.

Figura 9: Processo de Design Science Research Methodology (DSRM).



Fonte: Bordin (2015)

A tarefa de avaliação dos artefatos desenvolvidos a partir da DSRM é de grande importância. Lacerda *et al.* (2013) destacam que, para aumentar a confiabilidade nos resultados da pesquisa, é necessário um conjunto de cuidados e procedimentos especiais para o processo de avaliação e apresentam alguns métodos avaliação. O Quadro 7 apresenta alguns métodos de avaliação em *Design Science*.

Quadro 7: Métodos de avaliação em *Design Science*.

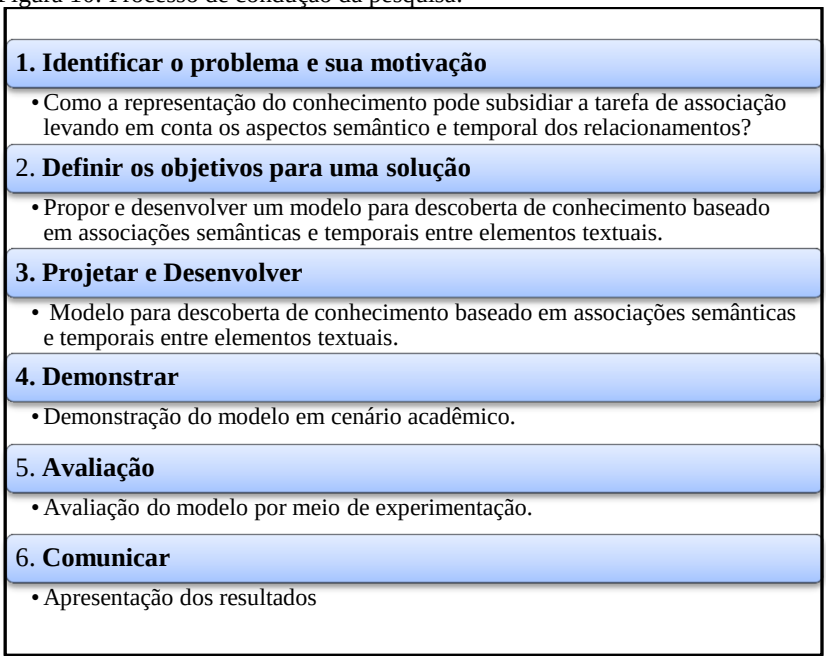
Forma de Avaliação	Métodos propostos
Observacional	Estudo de Caso: Estudar o artefato existente, ou não, em profundidade no ambiente de negócios. Estudo de Campo: Monitorar o uso do artefato em projetos múltiplos. Esses estudos podem, inclusive, fornecer uma avaliação mais ampla do funcionamento dos artefatos configurando, dessa forma, um método misto de condução da pesquisa.
Analítico	Análise Estatística: Examinar a estrutura do artefato para qualidades estáticas. Análise da Arquitetura: Estudar o encaixe do artefato na arquitetura técnica do sistema técnico geral. Otimização: Demonstrar as propriedades ótimas inerentes ao artefato ou então demonstrar os limites de otimização no comportamento do artefato. Análise Dinâmica: Estudar o artefato durante o uso para avaliar suas qualidades dinâmicas (por exemplo, desempenho).
Experimental	Experimento Controlado: Estudar o artefato em um ambiente controlado para verificar suas qualidades (por exemplo, usabilidade). Simulação: Executar o artefato com dados artificiais.
Teste	Teste Funcional (*Black Box*): Executar as interfaces do artefato para descobrir possíveis falhas e identificar defeitos. Teste Estrutural (*White Box*): Realizar testes de cobertura de algumas métricas para implementação do artefato (por exemplo, caminhos para a execução).
Descritivo	Argumento informado: Utilizar a informação das bases de conhecimento (por exemplo, das pesquisas relevantes) para construir um argumento convincente a respeito da utilidade do artefato. Cenários: Construir cenários detalhados em torno do artefato, para demonstrar sua utilidade.

Fonte: Lacerda *et al.* (2013)

3.3 PROCESSO DE CONDUÇÃO DA PESQUISA

A condução da presente pesquisa foi baseada nas etapas da DSRM, discutida na seção anterior, e está sintetizada na Figura 10.

Figura 10: Processo de condução da pesquisa.



Fonte: Autor.

Inicialmente, o processo de condução da pesquisa contou com uma revisão bibliométrica e analítica da literatura sobre o tema macro inicialmente proposto para esta tese: a descoberta de conhecimento em textos biomédicos (WOSZEZENKI; GONÇALVES, 2013). A revisão apontou diversas atividades para as quais os pesquisadores da mineração de textos biomédicos dedicam sua atenção: recuperação de documentos, classificação de documentos, extração de termos, anotação de documentos, extração de relacionamentos, mineração de imagens e desenvolvimento de ontologias. O estudo constatou que, dentre essas atividades, a que recebe maior destaque é a extração de relacionamentos explícitos ou implícitos (estes também chamados de associações), com quase metade das publicações. Isso se justifica pelo fato de que esta é

uma atividade fim da mineração de textos biomédicos, pois é ela que proporciona a descoberta de novas conexões, novos diagnósticos e tratamentos.

A revisão bibliométrica realizada permitiu constatar também que, embora muitos esforços tenham sido concentrados na extração de relacionamentos, existe um caminho longo a percorrer para que descobertas relevantes sejam realizadas, principalmente no que diz respeito à carga semântica e os aspectos temporais dos relacionamentos. Assim, o problema da pesquisa foi identificado (1) e o objetivo da tese foi estabelecido (2).

A terceira etapa da condução do processo refere-se ao Projeto e Desenvolvimento dos artefatos. Uma ontologia de alto nível foi criada para a representação semântica e temporal dos relacionamentos entre termos. Esta ontologia é detalhada no Capítulo 4 deste trabalho e compõe o modelo de descoberta do conhecimento desenvolvido e detalhado no Capítulo 5.

Na quarta etapa os artefatos (ontologia de relacionamentos e modelo) foram demonstrados em um cenário acadêmico, explorando os relacionamentos presentes no currículo Lattes. Essa demonstração é apresentada no Capítulo 5.

Quanto à avaliação do artefato, o método aqui empregado é, de acordo com a classificação de Lacerda et al. (2013), experimental, pois busca estudar o artefato em um ambiente controlado para verificar suas qualidades. Neste caso, o modelo proposto foi aplicado na solução de um problema clássico de Aprendizagem de Máquina e Mineração de textos, a classificação, com o intuito de verificar a sua capacidade de generalização/adaptabilidade e como as associações semânticas podem contribuir para a classificação de textos. O modelo proposto também é explorado para constatar que o aumento da força associativa entre termos ao longo do tempo pode indicar uma conexão direta desses termos no futuro. Os procedimentos de avaliação da pesquisa são detalhados no Capítulo 5.

4 MODELO PARA DESCOBERTA DE CONHECIMENTO

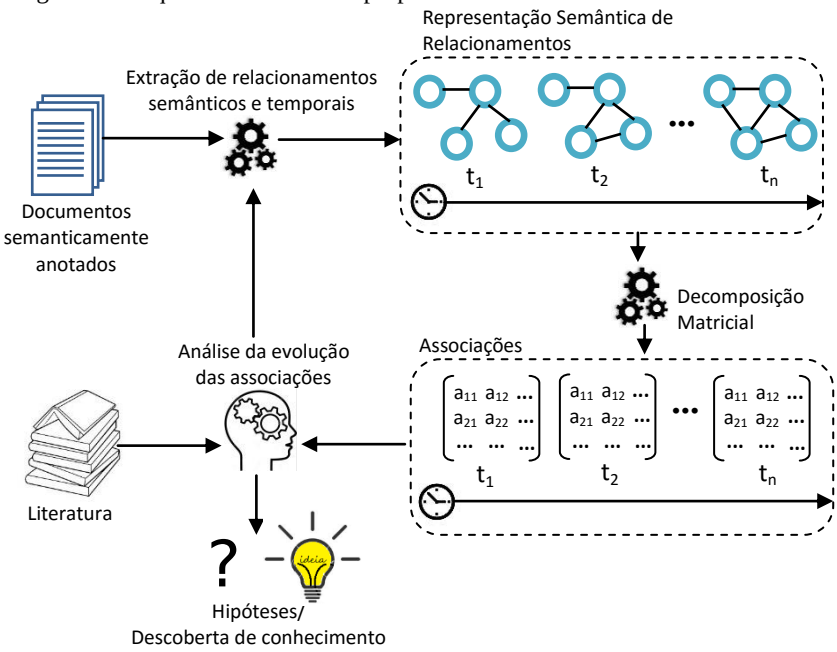
Para o entendimento do modelo proposto, inicialmente é apresentada a sua arquitetura, ilustrando seus componentes, as relações entre eles e a análise de cada etapa envolvida.

Em seguida, é demonstrado o funcionamento do modelo por meio de sua aplicação em um cenário fictício do contexto acadêmico.

4.1 MODELO PROPOSTO

A arquitetura do modelo proposto é ilustrada na Figura 11 e na sequência as etapas e componentes envolvidos são descritos.

Figura 11: Arquitetura do modelo proposto.



Fonte: Autor.

Documentos semanticamente anotados

No modelo proposto, a descoberta de conhecimento é realizada a partir de fontes textuais. Em fontes textuais a informação é não-estruturada e, por isso, os documentos devem estar anotados semanticamente, identificando as entidades, seus relacionamentos e a dimensão temporal desses relacionamentos. Uma vez que o modelo proposto é genérico, as fontes textuais podem pertencer a qualquer domínio do conhecimento.

É importante ressaltar que não faz parte do escopo desse trabalho a tarefa de anotação semântica dos documentos. Para efeito de avaliação do modelo proposto, considera-se um corpus já anotado.

Extração de relacionamentos semânticos e temporais

A partir dos documentos semanticamente anotados, ocorre a extração dos relacionamentos e suas informações semânticas e temporais. De forma mais detalhada, nessa fase necessitam ser extraídas as seguintes informações dos documentos semanticamente anotados:

- Entidades textuais (termos);
- Relacionamentos entre as entidades;
- Natureza (descrição) dos relacionamentos;
- Peso/força dos relacionamentos;
- Período de duração (tempo) dos relacionamentos;
- Propriedades dos relacionamentos;
- Peso/força das propriedades dos relacionamentos.

Representação Semântica de Relacionamentos

O modelo conta com representação semântica de relacionamentos para a população dos dados obtidos na etapa anterior. Nesse caso, a abordagem utilizada é ontologia.

Os dados extraídos dos documentos semanticamente anotados são utilizados para a população de uma ontologia de alto-nível para representação de relacionamentos. A ontologia permite representar os relacionamentos do domínio investigado, contendo todas as informações extraídas na etapa anterior. Esse artefato proporciona a representação do conhecimento com detalhamento semântico dos relacionamentos, além de representar as variações dos relacionamentos de acordo com o tempo (dimensão temporal). A apresentação detalhada sobre a ontologia de alto-nível criada neste trabalho acontece na Seção 4.2

Decomposição Matricial

A ontologia populada na etapa anterior é entrada para a decomposição matricial dos dados de forma a determinar a força de associação entre os termos. A LSI foi a técnica escolhida e a justificativa para isso é apresentada a seguir.

As técnicas de associação baseadas em grafos, regras de associação e SRI, embora possuam a capacidade de extrair associações entre termos, não são consideradas adequadas para a consecução do objetivo deste trabalho. Regras de associação não tratam da natureza das relações além de não conseguirem promover conexões latentes entre as conexões existentes. A SRI é uma técnica cujo funcionamento se assemelha à LSI e preocupa-se com o ônus computacional. Contudo, ela não é amplamente empregada e analisada se comparada à LSI. Além disso, a eficiência computacional não é uma das preocupações da presente pesquisa. Os grafos apresentam-se como uma forma de representação visual dos relacionamentos (diretos ou indiretos), mas os algoritmos de análise de redes não permitem identificar as contribuições indiretas dado um ponto em particular, como acontece na LSI.

A escolha pela LSI como técnica de associação entre termos neste trabalho justifica-se pelo seu alto poder em extrair relacionamentos e indiretos dentro de um conjunto de dados.

A ampla utilização da LSI por diversos pesquisadores para a descoberta de conexões latentes em bases textuais, conforme discutido na Seção 2.2.3.1, demonstra o sucesso desta técnica.

A entrada para esse processo é a ontologia de relacionamentos criada na etapa anterior. Nessa fase, as relações entre as instâncias da classe Entidade passam a ser representadas por uma matriz termo-termo, contendo todas as instâncias em questão. Cada entrada da matriz corresponde ao peso do relacionamento entre o par de instâncias.

Associações

Após a decomposição matricial, têm-se uma matriz dimensionalmente reduzida representando o grau de associação entre os termos. Nesta fase, pode-se identificar a força associativa entre entidades que não se relacionam diretamente, mas que podem exercer

influência uma sobre a outra por relacionarem-se diretamente com entidades em comum.

Uma vez que o modelo busca revelar a força das associações de acordo com o tempo, essa etapa gera diversas matrizes, uma para cada período de tempo expresso na ontologia. Assim, a evolução da força associativa pode ser avaliada pelo especialista.

Análise da evolução das associações

De posse das matrizes contendo o grau de associação entre os termos de acordo com o tempo, um especialista com conhecimento de domínio poderá analisar a evolução das associações de forma a: (i) identificar novos relacionamentos diretos; ou ii) levantar hipóteses.

Novos relacionamentos diretos podem ser constatados mediante a análise da literatura do domínio sob investigação. Ou seja, se o especialista identificar um crescimento na força da associação entre dois termos, após consultar a literatura, ele pode constatar um relacionamento direto entre esses termos. Nesse caso, a ontologia pode ser retroalimentada de forma a conter o novo relacionamento encontrado.

Caso não seja constatada a existência de um relacionamento direto entre dois termos cuja força de associação cresce ao passar do tempo, o especialista pode levantar hipóteses envolvendo esses dois termos.

Literatura

No modelo proposto, a literatura do domínio sob investigação pode ser analisada para a confirmação de supostos relacionamentos diretos levantados a partir da análise das associações.

4.2 ONTOLOGIA DE ALTO-NÍVEL PARA REPRESENTAÇÃO DE RELACIONAMENTOS

A noção de representação do conhecimento adotada neste trabalho é fortemente baseada em Sowa (2000), o qual trata a representação do conhecimento como um tema multidisciplinar que aplica teoria e técnicas de três outras áreas: *Lógica*, que fornece a estrutura formal e as regras de inferência; *Ontologia*, que define os tipos de coisas que existem no domínio abordado; e *Computação*, responsável

por promover suporte às aplicações que diferenciam a representação do conhecimento da filosofia pura. O autor argumenta que sem a lógica, a representação do conhecimento é vaga, sem critérios para determinar se as sentenças são redundantes ou contraditórias; sem ontologia, os termos e símbolos são mal definidos e confusos; e sem modelos computacionais, a lógica e a ontologia não podem ser implementadas em programas de computador. Assim, o autor afirma que a representação do conhecimento é a aplicação da lógica e da ontologia para a tarefa de construir modelos computáveis para algum domínio.

Neste trabalho, a ontologia foi escolhida como artefato para a representação do conhecimento por demonstrar ser a abordagem mais promissora, devido às suas características de permitir o compartilhamento de conhecimento entre humanos e entre agentes de *software*, bem como a reutilização de conhecimento entre aplicações. Além disso, ontologias podem ser utilizadas por motores de inferência para inferir novos conhecimentos a partir do conhecimento existente explicitado.

Além das justificativas encontradas na literatura, optou-se por essa abordagem, pois em uma ontologia, as relações podem ser semanticamente analisadas e entendidas, característica essa indispensável para a proposta dessa tese. Embora as ontologias de domínio sejam amplamente utilizadas pelos trabalhos que tratam da extração de relacionamentos e associações no domínio biomédico, não são encontradas ontologias de alto nível capazes de representar com riqueza de detalhes a semântica dos relacionamentos contidos nos textos.

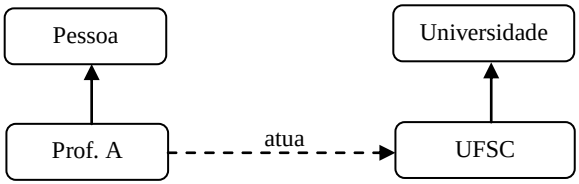
Levando em conta a importância dos aspectos semântico e temporal dos relacionamentos entre termos, bem como o potencial de representação oferecido pelas ontologias, esta seção apresenta uma ontologia de alto nível capaz de representar os detalhes semânticos e a dimensão tempo de um relacionamento.

Nesta seção, o termo “entidade” representa um termo ou elemento textual. Isso porque, no contexto da ontologia, um termo, ao ser associado a uma determinada classe, é entendido como uma entidade.

Considere o relacionamento entre um professor (Prof. A) e a universidade na qual ele atua (UFSC). A Figura 12 mostra a representação ontológica tradicional desse relacionamento por meio da propriedade “atua”. Como pode ser observado, o nível de detalhamento semântico deste relacionamento é pobre: não se conhece o período de atuação do professor na universidade e nem as características da

atuação. Em outras palavras o relacionamento se resume a apenas dois nós e uma aresta rotulada.

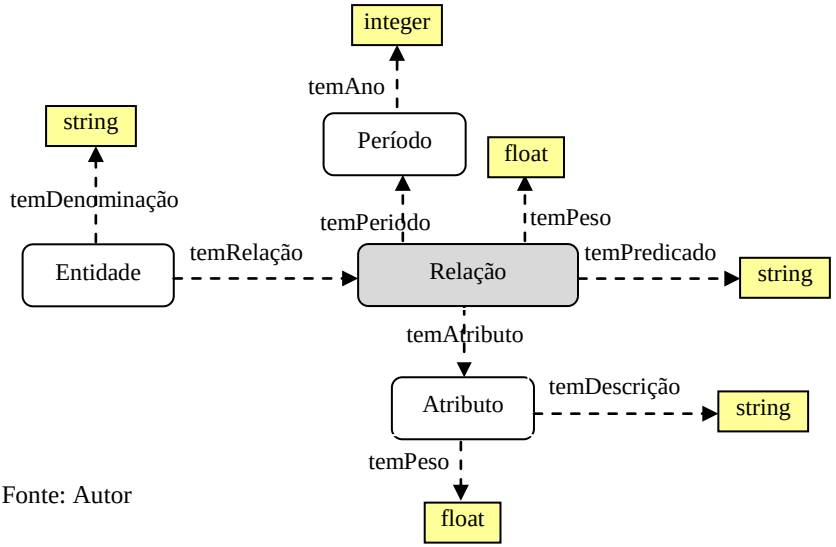
Figura 12: Representação ontológica tradicional de um relacionamento.



Fonte: Autor

Para aumentar o nível de detalhe semântico dos relacionamentos, a ontologia de alto nível proposta busca reificar o relacionamento entre termos, ou seja, ela introduz uma classe de relacionamento que conecta o sujeito ao objeto de forma que atributos adicionais (detalhes semânticos) sejam anexados a essa classe. Assim, ao contrário da representação tradicional das ontologias, onde o relacionamento entre entidades é apenas uma propriedade que conecta sujeito e objeto, na ontologia proposta, o relacionamento passa a ser uma classe com capacidade de conter diversas informações semânticas e temporais. A Figura 13 apresenta a ontologia de alto nível proposta.

Figura 13: Ontologia de alto nível para representação de relacionamentos.



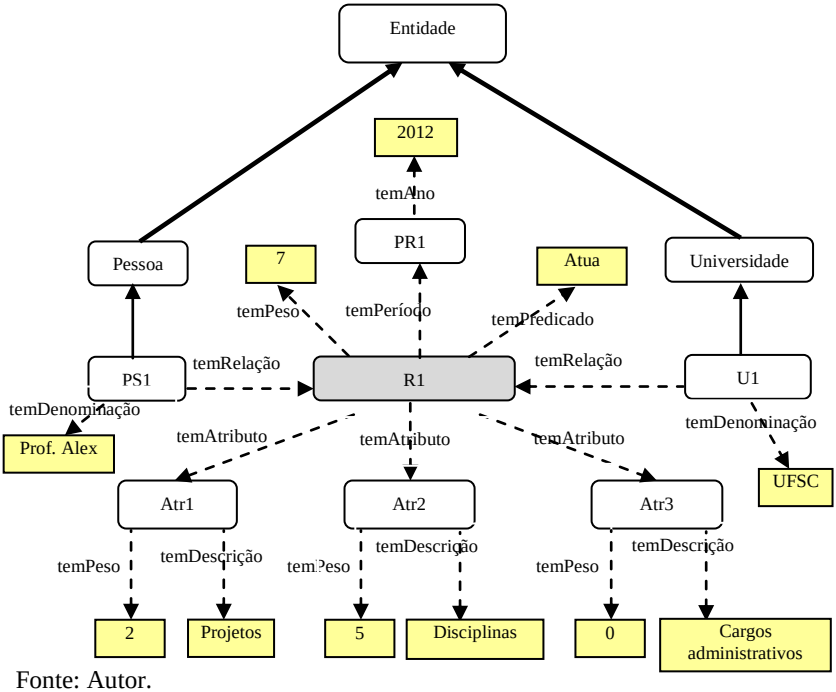
Fonte: Autor

Na ontologia proposta, toda entidade conecta-se a outra entidade por meio da classe **Relação**, a qual possui propriedades consideradas importantes para expressar características semânticas de um relacionamento, a saber:

- **temPredicado**: O predicado expressa a natureza de um relacionamento, denotando a semântica do mesmo. Por exemplo, na sentença "O gene G inibe a doença D", o relacionamento entre G e D possui predicado "inibe";
- **temPeriodo**: A temporalidade do relacionamento é expressa por meio da propriedade "temPeriodo" que indica o período em que o relacionamento acontece. Para fins de simplificação, a classe Período possui apenas a propriedade "temAno", representando períodos anuais de tempo. Contudo, caso necessário, pode-se variar a granularidade da dimensão tempo de forma a representar mês, dia, semana, entre outros;
- **temPeso**: Força pela qual duas entidades estão conectadas, determinando a relevância do relacionamento. O peso de um relacionamento pode variar de acordo com o tempo indicando se ele está se enfraquecendo ou se fortalecendo. Na ontologia ele é representado por meio da propriedade "temPeso";
- **temAtributo**: Um relacionamento pode ter diversos atributos que representam características importantes capazes de contribuir para o enriquecimento semântico do mesmo. A propriedade "temAtributo" conecta a classe Relação à classe Atributo. Todo atributo possui uma descrição (propriedade "temDescrição") e um peso associado (propriedade "temPeso"). Os pesos dos atributos do relacionamento podem ser utilizados para calcular o peso do relacionamento.

Conforme visto na seção 2.4.1, uma ontologia de alto-nível pode ser especializada para a criação de uma ontologia de domínio. Assim, a ontologia proposta pode ser estendida para melhor representar o domínio que se deseja modelar. No caso do exemplo inicial (relacionamento de um professor com uma universidade), pode-se criar duas subclasses de Entidade: Pessoa e Universidade. A Figura 14 mostra uma possível representação do relacionamento entre o professor e a universidade utilizando a ontologia proposta.

Figura 14: Representação ontológica de um relacionamento utilizando a ontologia proposta. → Subclasse; → Instância; - ► Propriedade



Fonte: Autor.

Na representação utilizando a ontologia proposta, o Prof. A, representado pela entidade PS1 da classe “Pessoa” conecta-se com a UFSC, representada pela entidade U1 da classe “Universidade” por meio da entidade R1 da classe Relação. O relacionamento representado acontece no período de 2012 e leva em conta atributos importantes para o domínio, como os projetos executados, as disciplinas ministradas e os cargos administrativos exercidos. O peso do relacionamento (determinado pelo valor 7) considera a soma dos pesos de cada atributo.

O relacionamento do professor com a universidade pode ser representado também em anos anteriores ao de 2012, apresentando pesos diferentes para cada atributo e, consequentemente, para o relacionamento. Com isso, é possível conhecer a evolução do relacionamento entre o professor e a universidade ao longo do tempo.

Em uma rede de entidades conectadas diretamente, a evolução da força dos relacionamentos ao longo do tempo pode ajudar a descobrir

conexões implícitas (associações) relevantes de forma a prever futuros relacionamentos diretos.

4.3 DEMONSTRAÇÃO DA APLICABILIDADE DA ONTOLOGIA

De forma a demonstrar a aplicabilidade e o potencial de uso da ontologia proposta, será explorado o cenário acadêmico e profissional, analisando os relacionamentos presentes no currículo Lattes.

As informações presentes no currículo Lattes revelam relacionamentos importantes estabelecidos durante a vida acadêmica de um indivíduo, como relacionamentos com instituições, pessoas, áreas do conhecimento e linhas de pesquisa. O tipo de relacionamento estabelecido também pode ser identificado, ou seja, pode-se saber se o indivíduo estudou e/ou atuou em uma determinada instituição, se ele é orientador de pesquisa ou colega de trabalho de uma determinada pessoa ou se tem apenas um relacionamento de coautoria com essa pessoa. Ainda, o Currículo Lattes oferece diversos atributos que podem enriquecer as informações de um relacionamento. Considere, por exemplo, o número de projetos de pesquisa executados, disciplinas ministradas e cargos administrativos assumidos pelo indivíduo durante o período em que teve um relacionamento de atuação com uma organização. Esses atributos descrevem o relacionamento e podem ajudar a conferir uma força ao mesmo, determinando sua relevância.

Além disso, o Currículo Lattes, armazena a informação temporal dos relacionamentos, ou seja, o período de tempo em que os relacionamentos ocorrem. Assim, é possível avaliar a evolução histórica das relações de um indivíduo com outras entidades.

O currículo escolhido para esta demonstração foi o de um dos coautores deste trabalho, o professor Alexandre Leopoldo Gonçalves, professor da Universidade Federal de Santa Catarina (UFSC). Para uma melhor visualização, os relacionamentos são representados em forma de tabelas. Além disso, por razões de espaço e em função do nível de detalhes, foi escolhida apenas a principal instituição de atuação do professor, a UFSC. O Quadro 8 apresenta a evolução do relacionamento do professor Alexandre com a UFSC desde o ano em que ele iniciou sua atuação profissional (2010) até o momento da coleta dos dados na plataforma Lattes (31 de agosto de 2013). Todas as propriedades da ontologia proposta são representadas nesta tabela: predicado (tipo do relacionamento), peso do relacionamento, atributos, peso de cada atributo e período.

Quadro 8: Relacionamentos do Professor Alexandre Leopoldo Gonçalves com a UFSC.

	2010		2011		2012		2013	
	Predicado: Professor Adjunto		Predicado: Professor Adjunto		Predicado: Professor Adjunto		Predicado: Professor Adjunto	
	Peso*: 7		Peso*: 13		Peso*: 14		Peso*: 15	
	Atributos	Peso	Atributos	Peso	Atributos	Peso	Atributos	Peso
UFSC	Disciplinas	4	Disciplinas	6	Disciplinas	5	Disciplinas	7
	Pesquisa	1	Pesquisa	1	Pesquisa	2	Pesquisa	2
	Extensão	1	Extensão	2	Extensão	2	Extensão	1
	Cargos Administrat.	1	Cargos Administrat.	4	Cargos Administrat.	5	Cargos Administrat.	5

* Para o cálculo do peso do relacionamento foi realizada a soma aritmética dos pesos de cada atributo. Contudo, outras fórmulas podem ser empregadas.
Fonte: Autor

As informações extraídas do currículo Lattes mostram que o Alexandre tem uma relação de atuação como Professor Adjunto (Predicado) com a UFSC. O relacionamento é descrito em termos dos atributos: disciplinas, pesquisa, extensão e cargos administrativos. No período inicial de sua atuação (2010), o professor Alexandre ministrou 4 disciplinas, executou 1 projeto de pesquisa, 1 projeto de extensão e assumiu 1 cargo administrativo, conferindo peso 7 ao seu relacionamento com a UFSC nesse período. No período seguinte (2011) a força da relação do professor Alexandre com a UFSC aumenta consideravelmente (peso 13), uma vez que ele ministrou mais disciplinas, executou mais projetos de extensão e assumiu mais cargos administrativos do que no período anterior. Percebe-se ainda, que a força do relacionamento segue aumentando nos períodos subsequentes na medida em que os pesos dos atributos considerados variam.

A ontologia proposta confere flexibilidade na escolha dos atributos que descrevem um relacionamento, ou seja, esses atributos podem variar de acordo com a necessidade do contexto que está sendo modelado. No exemplo aqui explorado, o nível de detalhes do relacionamento pode ser aumentado, como mostra o Quadro 9. Novos atributos surgem para descrever o relacionamento, como Professor do PPGEGC (Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento), Membro do NDE (Núcleo Docente Estruturante), Membro de Colegiado de Curso, Coordenação de Curso e Subcoordenação de Curso. Os últimos 4 atributos citados no Quadro 8 estão reunidos no atributo Cargos Administrativos do Quadro 9.

Quadro 9: Relacionamentos do Professor Alexandre Leopoldo Gonçalves com a UFSC em maior nível de detalhes.

	2010		2011		2012		2013	
UFSC	Predicado: Professor Adjunto		Predicado: Professor Adjunto		Predicado: Professor Adjunto		Predicado: Professor Adjunto	
	Peso: 7		Peso: 13		Peso: 14		Peso: 15	
	Atributos	Peso	Atributos	Peso	Atributos	Peso	Atributos	Peso
	Disciplinas	4	Disciplinas	6	Disciplinas	5	Disciplinas	7
	Professor PPGEGC	1	Professor PPGEGC	1	Professor PPGEGC	1	Professor PPGEGC	1
	Subcoordenação de curso	1	Membro do NDE	2	Membro do NDE	2	Membro do NDE	2
	Pesquisa	1	Membro de Colegiado de Curso	2	Membro de Colegiado de Curso	2	Membro de Colegiado de Curso	2
	Extensão	1	Pesquisa	1	Coord. de curso	1	Coord. de curso	1
			Extensão	2	Pesquisa	2	Pesquisa	2
					Extensão	2	Extensão	1

Fonte: Autor.

Outras informações relevantes presentes no Currículo Lattes são as áreas do conhecimento, linhas de pesquisa e palavras-chave atacadas pelo pesquisador. O Quadro 10 apresenta a evolução dos relacionamentos do professor Alexandre com as 8 palavras-chave que mais ocorrem em seu Lattes. Na tabela estão representados os períodos que vão de 1998 até o momento da coleta dos dados e o peso de cada relacionamento, neste caso, o número de ocorrências da palavra-chave no período.

De forma a ilustrar apenas a variação da força dos relacionamentos entre o professor Alexandre e os termos de interesse, o predicado e os atributos foram suprimidos nesta tabela. Contudo, uma vez que o predicado deve indicar o tipo de relacionamento existente, neste caso, pode-se usar o termo “publica” para descrevê-los. Ainda, os atributos que descrevem os relacionamentos podem ser definidos em termos dos tipos de publicações realizadas: artigos em periódicos, artigos em anais, capítulo de livro, entre outros.

Quadro 10: Peso dos relacionamentos do Professor Alexandre Leopoldo Gonçalves com as 8 palavras-chave que mais ocorrem em seu currículo Lattes.

Palavra-Chave	**Total**	**2013**	**2012**	**2011**	**2010**	**2009**	**2008**	**2007**	**2006**	**2005**	**2004**	**2003**	**2002**	**2001**	**2000**	**1999**	**1998**
Engenharia do Conhecimento	**17**		1	4	3	4	1	1	3								
Mineração de Textos	**14**	1	2			1		1	2	7							
Gestão do Conhecimento	**11**			1	1		1	1	7								
Recuperação de Informação	**11**	2				1	1	2	1		2		1	1			
Plataforma Lattes	**9**												1		3	5	
Mineração de Dados	**8**					1							2	2	1	2	
Data Warehouse	**7**												2	2	1	1	1
Ontologias	**7**		2	2	1	2											

Fonte: Autor

Analisar a variação da força dos relacionamentos permite conhecer a evolução do interesse do pesquisador por determinados temas de pesquisa, bem como estimar tendências para futuras pesquisas. No caso do professor Alexandre, pode-se dividir seu histórico de pesquisa em dois períodos. No primeiro período (1998 a 2002), seus trabalhos tiveram como principais focos a Plataforma Lattes, Mineração de Dados e *Data Warehouse*. No segundo período (2005 a 2013), seus estudos focam na Engenharia do Conhecimento, Mineração de Textos, Gestão do Conhecimento e Ontologias. Já a Recuperação de Informação é um tema de interesse em ambos os períodos.

A Engenharia do Conhecimento se destaca como principal tema de interesse do professor Alexandre. Ainda, os trabalhos com Ontologias, uma ferramenta da Engenharia do Conhecimento, fortalecem ainda mais o relacionamento do pesquisador com esse tema.

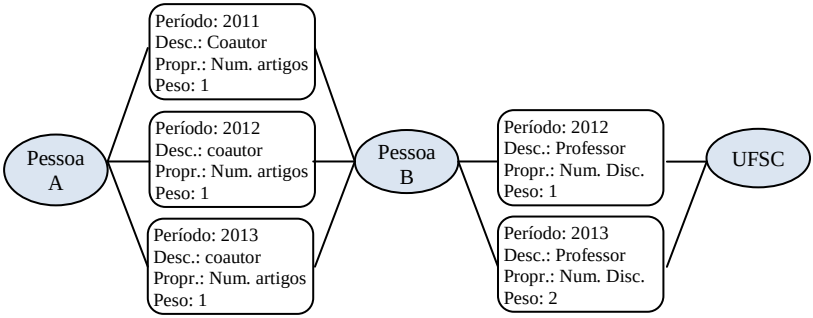
Nessa seção, a ilustração da aplicabilidade da ontologia proposta demonstra suas principais contribuições:

- Alto poder de representação uma vez que pode ser estendida para outros domínios;
- Grande poder de expressividade da semântica dos relacionamentos;
- Flexibilidade no estabelecimento dos atributos e respectivos pesos (podem ser estabelecidos de acordo com a necessidade e interesse do usuário/domínio);
- Representação da dimensão temporal dos relacionamentos, mantendo um histórico da evolução das conexões.

4.4 DEMONSTRAÇÃO DA APLICABILIDADE DO MODELO

Nessa seção é apresentado um cenário fictício do contexto acadêmico, no qual o modelo proposto é aplicado de forma a demonstrar o seu funcionamento. Os relacionamentos pertinentes a esse cenário são representados pela ontologia de alto nível proposta nessa tese. A Figura 15 ilustra a ontologia contendo a representação do conhecimento do cenário em questão. Para efeitos de simplificação da representação gráfica da ontologia, as características dos relacionamentos (descrição, peso do relacionamento, propriedades e seus pesos e período) estão todas agrupadas.

Figura 15: Ontologia de um cenário fictício no contexto acadêmico.



Fonte: Autor

Pessoa A e Pessoa B são entidades do tipo Pessoa, e UFSC é uma entidade do tipo Universidade. As pessoas A e B publicam artigos em parceria e, por isso, têm relacionamento do tipo “coautor”; já a Pessoa B é professora da UFSC. O relacionamento de coautoria entre as pessoas A e B é descrito em termos do número de artigos que produziram em conjunto e acontece nos períodos 2011, 2012 e 2013. O relacionamento da Pessoa B com a UFSC acontece nos períodos 2012 e 2013 e é descrito em termos do número de disciplinas que ela ministra.

Para aplicação da técnica LSI, a ontologia deve ser transformada para o formato matricial cujas linhas e colunas contêm todos os termos (entidades) que estão representados na ontologia. Cada entrada da matriz corresponde ao peso do relacionamento entre o par de entidades (zero indica a inexistência de relacionamento). Nessa etapa,

uma matriz deverá ser gerada para cada período da ontologia. Assim, as matrizes termo-termo para a ontologia da Figura 15 são apresentadas no Quadro 11.

Quadro 11: Representação matricial dos relacionamentos presentes na ontologia.

2011

	P. A	P. B	UFSC
P. A	0	1	0
P. B	1	0	0
UFSC	0	0	0

2012

	P. A	P. B	UFSC
P. A	0	1	0
P. B	1	0	1
UFSC	0	1	0

2013

	P. A	P. B	UFSC
P. A	0	1	0
P. B	1	0	2
UFSC	0	2	0

Fonte: Autor

A próxima etapa do modelo diz respeito à aplicação da técnica LSI sobre as matrizes criadas a partir da ontologia. Essa técnica, conforme explicada na seção 2.2.3, decompõe uma matriz em outras três matrizes e realiza a redução de dimensionalidade através da utilização dos *k* fatores (valores singulares) mais relevantes, nesse exemplo, *k*=2. As tabelas do Quadro 12 demonstram o resultado após a aplicação da LSI. Os resultados percebidos são analisados na sequência.

Quadro 12: Matrizes resultantes do cálculo da LSI, com k = 2.

2011

	P. A	P. B	UFSC
P. A	1,000	0,000	**-0,000**
P. B	0,000	1,000	0,000
UFSC	-0,000	0,000	0,000

2012

	P. A	P. B	UFSC
P. A	0,707	-0,000	**0,707**
P. B	-0,000	1,414	-0,000
UFSC	0,707	-0,000	0,707

2013

	P. A	P. B	UFSC
P. A	0,447	0,000	**0,894**
P. B	0,000	2,236	-0,000
UFSC	0,894	-0,000	1,789

Fonte: Autor

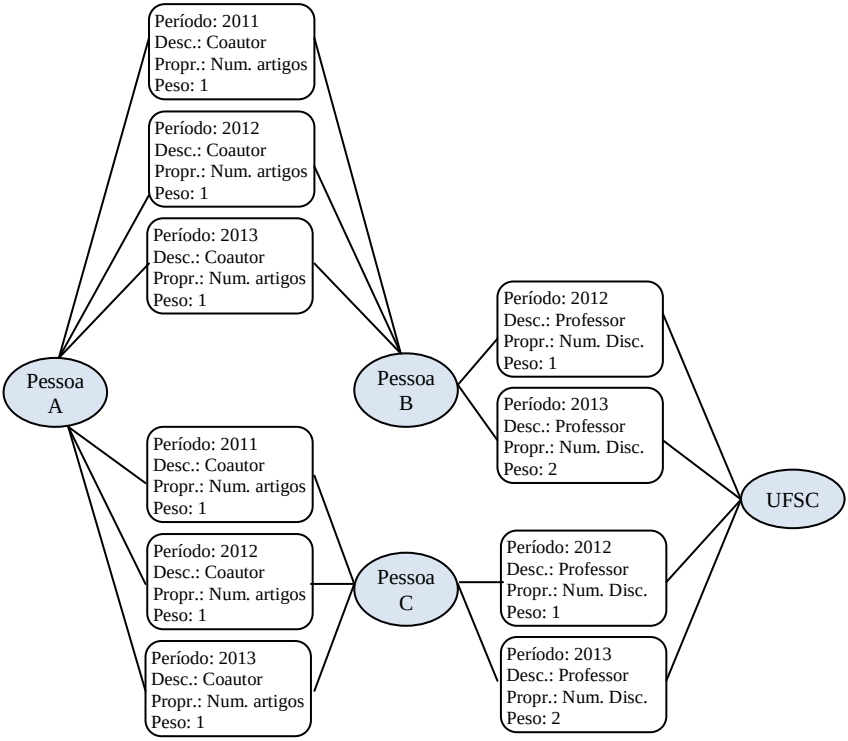
Na ontologia que representa o cenário avaliado, as pessoas A e B possuem relacionamento com peso 1 em todos os períodos. O relacionamento da Pessoa B com a UFSC possui peso 1 em 2012 e se fortalece em 2013, passando a ter peso 2. Embora a Pessoa A não possua relacionamento com a UFSC em nenhum dos períodos, um grau de associação é atribuído nesse caso em função do relacionamento indireto via Pessoa B. Percebe-se que o fortalecimento do relacionamento da Pessoa B com a UFSC fortalece também a associação entre a Pessoa A e a UFSC, ou seja, o grau de associação da Pessoa A com a UFSC passa de 0,707 em 2012 para 0,894 em 2013. Assim, quanto maior o peso dos relacionamentos maior o grau de associação entre as entidades envolvidas nesses relacionamentos.

A diagonal principal das matrizes resultantes representa a contribuição de cada um dos termos na distribuição dos pesos das associações indiretas. Na medida em que o peso do(s) relacionamento(s) de um determinado termo aumenta, o potencial de contribuição desse termo para as associações também aumenta.

Observa-se que a biblioteca utilizada nessa implementação retira da matriz resultante o peso dos relacionamentos diretos. Contudo, caso necessário, eles podem ser reconstituídos na matriz resultante.

A ontologia da Figura 15 foi acrescentada mais uma entidade, Pessoa C, da classe Pessoa. A Pessoa C relaciona-se com a Pessoa A (relacionamento de coautoria) e com a UFSC (professor). O novo cenário é apresentado na ontologia da Figura 16.

Figura 16: Ontologia após a inserção da entidade Pessoa C.



Fonte: Autor.

As matrizes termo-termo para a ontologia da

Figura 16 são apresentadas no Quadro 13.

Quadro 13: Representação matricial dos relacionamentos presentes na ontologia após a inserção da entidade Pessoa C ao cenário.

2011

	P. A	P. B	P. C	UFSC
P. A	0	1	1	0
P. B	1	0	0	0
P. C	1	0	0	0
UFSC	0	0	0	0

2012

	P. A	P. B	P. C	UFSC
P. A	0	1	1	0
P. B	1	0	0	1
P. C	1	0	0	1
UFSC	0	1	1	0

2013

	P. A	P. B	P. C	UFSC
P. A	0	1	1	0
P. B	1	0	0	2
P. C	1	0	0	2
UFSC	0	2	2	0

Fonte: Autor.

As tabelas do Quadro 14 demonstram o resultado após a aplicação da LSI, com $k = 2$. Os resultados são observados a seguir.

Quadro 14: Matrizes resultantes do cálculo da LSI, após a inserção da Pessoa C, com $k = 2$.

2011

	P. A	P. B	P. C	UFSC
P. A	1,414	-0,000	-0,000	**-0,000**
P. B	-0,000	0,707	**0,707**	-0,000
P. C	-0,000	0,707	0,707	-0,000
UFSC	-0,000	-0,000	-0,000	0,000

2012

	P. A	P. B	P. C	UFSC
P. A	1,000	0,000	0,000	**1,000**
P. B	0,000	1,000	**1,000**	-0,000
P. C	0,000	1,000	1,000	-0,000
UFSC	1,000	-0,000	-0,000	1,000

2013

	P. A	P. B	P. C	UFSC
P. A	0,632	0,000	0,000	**1,265**
P. B	0,000	1,581	**1,581**	0,000
P. C	0,000	1,581	1,581	0,000
UFSC	1,265	0,000	0,000	2,530

Fonte: Autor.

No período de 2011, a Pessoa C relaciona-se apenas com a Pessoa A. Levando em conta que a Pessoa A tem um relacionamento com a Pessoa B, a Pessoa C e a Pessoa B possuem um relacionamento indireto, resultando num grau de associação de 0,707.

No período de 2012, a Pessoa C se relaciona com a Pessoa A com a UFSC. A Pessoa B também se relaciona com a Pessoa A e com a UFSC. Dessa forma, percebe-se que o grau de associação entre a Pessoa A e a UFSC aumenta (1,000), se comparado ao cenário anterior (0,707). Isso se deve ao fato de que nesse cenário a Pessoa A possui dois relacionamentos indiretos com a UFSC: um via Pessoa B e outro via Pessoa C (no cenário anterior, o relacionamento indireto se ocorria apenas via Pessoa B). O grau de associação da Pessoa C com a Pessoa B também aumenta (de 0,707 em 2011 para 1,000 em 2012) devido ao relacionamento indireto via Pessoa A e via UFSC.

O aumento da força dos relacionamentos no período de 2013 ocasiona o aumento na força associativa entre as pessoas B e C (1,581) e entre a Pessoa A e a UFSC (1,265).

Embora esteja sendo utilizada apenas uma única propriedade para descrever um relacionamento (“coautor” ou “professor”) no cenário

apresentado, vale lembrar que, conforme discutido no Capítulo 3, um relacionamento pode conter diversas propriedades com seus respectivos pesos. Assim, o modelo permite a flexibilização da descoberta de associações latentes uma vez que o usuário pode escolher quais propriedades são mais interessantes para a descoberta de conhecimento no domínio investigado.

4.5 DIFERENCIAL DO MODELO PROPOSTO

Alguns dos trabalhos relacionados fazem uso de ontologias biomédicas em suas abordagens de descoberta de associações. Nesses casos, as ontologias são utilizadas apenas para refinar os termos extraídos dos documentos, tornando-os equivalentes aos termos definidos na ontologia.

O modelo para descoberta de conhecimento proposto nesta pesquisa diferencia-se dos demais trabalhos pela utilização de ontologia para representação das relações aumentando o nível de detalhamento da semântica dos relacionamentos e possibilitando a representação temporal dos mesmos. A ontologia é de alto nível e facilmente aplicável para vários domínios.

Apenas três trabalhos foram identificados como fortemente relacionados a esta tese por tratarem, além da semântica, da dimensão temporal dos relacionamentos. No caso de Cohen e Schvanveldt (2010), eles utilizam a frequência da coocorrência dos termos, ou seja, o tipo (descrição) de relacionamento não é conhecido. No presente trabalho, os relacionamentos possuem propriedades (características semânticas), cada qual podendo ser ponderada de acordo com a necessidade do usuário. O grau de associação entre duas entidades é determinado pelo peso dos relacionamentos entre as entidades do domínio.

Embora Xu et al. (2013) considerem a semântica e a temporalidade das relações, o modelo desses autores não oferece nenhum mecanismo para medir a força das relações diretas e indiretas.

Já Bovo (2011) considera a temporalidade das relações, mede a força das relações diretas e indiretas e adiciona a dimensão “relação” ao tema de análise. Contudo, a “relação” denota apenas um único tipo de relacionamento expresso em uma ontologia de domínio. Em outras palavras, uma determinada matriz de correlação/associação temporal refere-se a apenas um único tipo de relacionamento. Diferentemente do trabalho de Bovo (2011), a ontologia integrante do modelo proposto nesta tese foca na representação semântica dos diversos relacionamentos entre entidades, permitindo que a matriz de relacionamentos e

consequentemente de associações contemple todos os relacionamentos representados na ontologia de domínio. Além disso, a ontologia permite que um relacionamento seja determinado por diversas características ou atributos, cada um possuindo um peso para o relacionamento. Outra diferença entre ambos os modelos é a técnica utilizada para derivar a matriz de associação a partir da matriz de relacionamentos (correlação no caso de Bovo (2011)). Bovo (2011) utiliza a similaridade entre os vetores de contexto de dois conceitos (*i*, *j*) enquanto que a presente pesquisa utiliza LSI.

5 AVALIAÇÃO

Este capítulo apresenta a avaliação do modelo realizada perante dois tipos de experimentos. O primeiro trata da associação semântica e temporal entre termos com o objetivo de constatar que o aumento da força associativa entre termos ao longo do tempo pode indicar uma conexão direta desses termos no futuro. O segundo experimento lida com a classificação de documentos, um problema clássico de Aprendizagem de Máquina e Mineração de textos, com o intuito de verificar a sua capacidade de generalização/adaptabilidade e como as associações podem contribuir para a classificação de textos.

Para cada um dos experimentos foi construído um protótipo. Os detalhes são apresentados a seguir.

5.1 ASSOCIAÇÃO

Os experimentos descritos nesta seção são baseados no trabalho relacionado de Cohen e Schvaneveldt (2010), que utilizam os contextos de duas descobertas seminais de Swanson: 1) a conexão do óleo de peixe com a Síndrome de Raynaud; e 2) a conexão do magnésio com a dor de cabeça para aplicar um método por eles criado, a SRI, com o mesmo objetivo de prever conexões diretas a partir de associações ao longo do tempo.

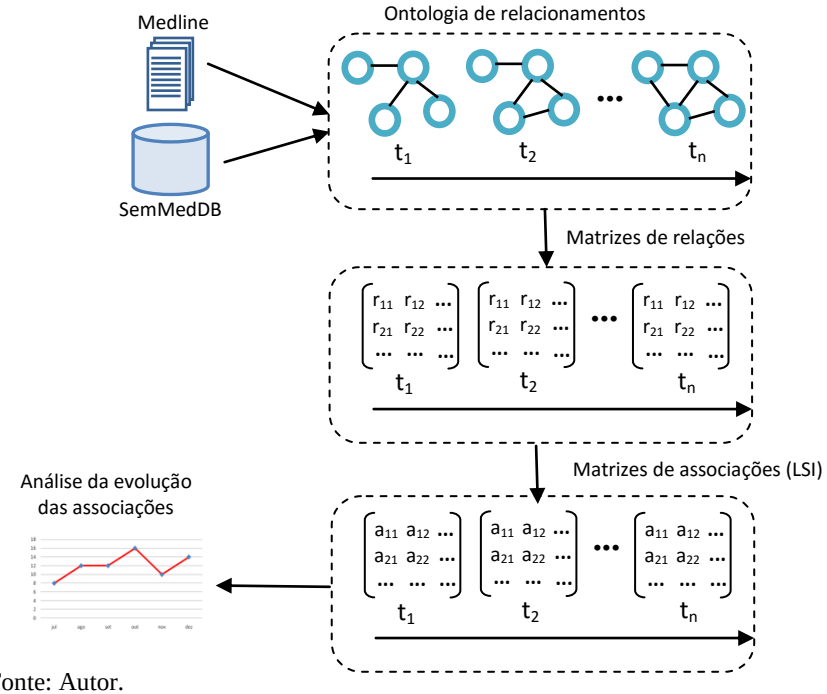
Além de replicar as duas descobertas acima citadas, este trabalho buscou verificar o comportamento do modelo proposto quando varia-se a semântica dos relacionamentos entre termos.

O protótipo para a realização dos experimentos de associação foi desenvolvido na linguagem de programação Java seguindo as fases do modelo proposto. Ele está ilustrado na Figura 17 e é detalhado a seguir.

As fontes de informação para a entrada do processo são o **Medline** e o **SemMedDB**. O Medline, já mencionado no Capítulo 1, é uma base de dados bibliográfica e disponibiliza mais de 22 milhões de citações e resumos de publicações científicas das áreas da medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular, entre outros. O SemMedDB (KILICOGLU et al., 2012) é uma base de dados relacional contendo predicações semânticas (sujeito-predicado-objeto) extraídas dos títulos e resumos do Medline. Ela também disponibiliza a informação temporal (ano) em que ocorre um relacionamento.

Do SemMedDB são extraídos os termos que farão parte da ontologia (explicado detalhadamente nas subseções 5.1.1 e 5.1.2) e do Medline são extraídos os pesos dos relacionamentos entre os termos baseados na coocorrência, para cada ano investigado. Com esses dados, a **ontologia de relacionamentos** é instanciada e carregada.

Figura 17: Protótipo desenvolvido para os experimentos de associação.



Fonte: Autor.

Posteriormente, a ontologia de relacionamentos instanciada é representada matricialmente, criando **matrizes de relacionamentos**. Cria-se uma matriz para cada ano que está representado na ontologia.

Na sequência, a LSI é aplicada sobre cada uma das matrizes, resultando nas **matrizes de associações** entre termos. A biblioteca utilizada para essa fase é a *Efficente Java Matrix Library*[6] (EJML), uma biblioteca desenvolvida em Java para a manipulação de matrizes densas.

A seguir são detalhados os dois experimentos que replicam as descobertas seminais de Swanson.

[6] Disponível em: http://ejml.org/wiki/index.php?title=Main_Page.

5.1.1 Experimento 1: Raynaud e Eicosapentaenoic

O primeiro experimento faz o levantamento da força associativa ao longo do tempo entre os termos “Raynaud”, representando a Síndrome de Raynaud, e “Eicosapentaenoic”, representando o ácido eicosapentaenoico, o ingrediente ativo do óleo de peixe (COHEN; SCHVANEVELDT, 2010).

De acordo com o protótipo ilustrado na Figura 17, inicialmente realiza-se a extração das informações presentes no Medline e no SemMedDB. Assim, a partir de consultas SQL ao SemMedDB, foram identificados os 24 vizinhos mais próximos do termo “Raynaud” (24 termos que mais se relacionam com o termo “Raynaud”) e os 24 vizinhos mais próximos do termo “Eicosapentanoic” (24 termos que mais se relacionam com o termo “Eicosapentaenoic”), além de ambos os termos, resultando no total de 50 termos.

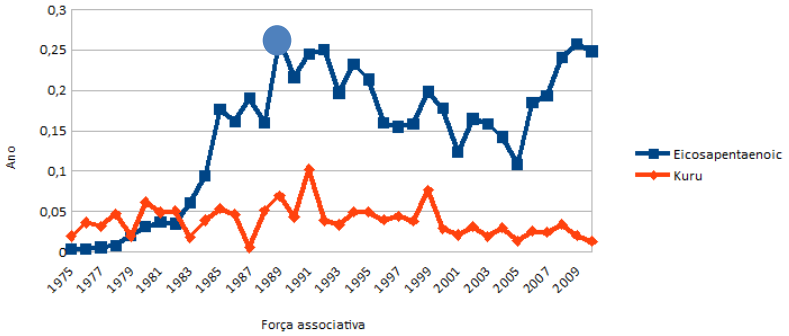
O relacionamento entre os termos possui predicado “*genérico*”, pois são considerados todos os predicados existentes entre dois termos, ou seja, são considerados todos os tipos de relacionamentos entre dois termos. Assim, o peso do relacionamento entre dois termos foi obtido a partir dos títulos e resumos de todas as citações do Medline entre 1975 e 2010, baseado na coocorrência.

De posse dos termos, relacionamentos e seus respectivos pesos e do tempo, a ontologia de relacionamentos foi instanciada. Em seguida, foi realizada a representação matricial da ontologia, gerando 35 matrizes (uma para cada ano entre 1975 a 2010) de tamanho 50x50.

Sob cada matriz de relações foi executado o processo da LSI de forma a calcular a força associativa entre os termos “Raynaud” e “Eicosapentaenoic”.

O gráfico da Figura 18 apresenta os resultados do experimento.

Figura 18: Evolução da força associativa entre os termos "Raynaud" e "Eicosapentaenoic" e os termos "Raynaud" e "Kuru" entre 1975 a 2010.



Fonte: Autor.

A curva de evolução associativa apresentada no gráfico se comporta de forma análoga ao experimento de Cohen e Schvaneveldt, conforme mostrado na Figura 7 da Seção 2.5.3. Assim, corroborando com Cohen e Schvaneveldt (2010), a figura mostra um aumento na força associativa entre "Raynaud" e "Eicosapentaenoic" que precede a primeira vez que esses termos coocorrem em um resumo do Medline, indicado por um círculo.

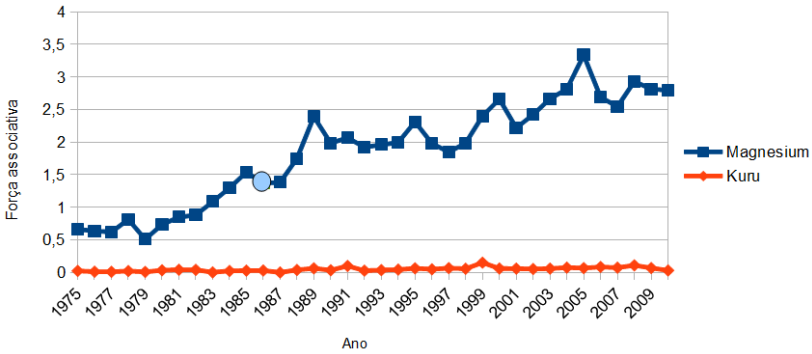
Observa-se também que a força associativa entre "Raynaud" e "Kuru" não apresenta crescimento significativo e esse padrão se mantém ao longo dos anos. Isso demonstra a diferença de comportamento entre as curvas de associação entre dois termos que possuem relação/associação na literatura e entre dois termos que não possuem relação/associação.

5.1.2 Experimento 2: Migraine e Magnesium

O segundo experimento faz o levantamento da força associativa ao longo do tempo entre os termos "Migraine" e "Magnesium", outra Descoberta Baseada na Literatura de Swanson (SWANSON, 1988).

O processo de instanciação do modelo segue os mesmos passos do experimento anterior. Neste caso, foram extraídos os 24 vizinhos mais próximos do termo "Migraine" e os 24 vizinhos mais próximos do termo "Magnesium". A Figura 19 apresenta os resultados desse experimento.

Figura 19: Evolução da força associativa entre os termos "Migraine" e "Magnesium" e os termos "Migraine" e "Kuru" entre 1975 a 2010.



Fonte: Autor.

O mesmo comportamento do experimento anterior é observado aqui. A curva de evolução associativa apresentada no gráfico deste experimento também se comporta de maneira simiar ao experimento de Cohen e Schvaneveldt, conforme mostrado na Figura 8 da Seção 2.5.3, ou seja, a força associativa entre “Migraine” e “Magnesium” aumenta anos antes da primeira coocorrência, demonstrada no gráfico pelo círculo azul, e mantém o padrão de crescimento ao longo o tempo.

5.1.3 Experimento 3: Variando a semântica dos relacionamentos

Os experimentos de associação acima apresentados levam em conta todos os tipos de relacionamentos existentes entre dois termos. Ou seja, o peso do relacionamento entre dois termos é a soma dos pesos de todos os tipos de relacionamentos existentes entre ambos os termos. Contudo, o modelo proposto permite especificar a semântica dos relacionamentos na representação do conhecimento de domínio. Em outras palavras, é possível determinar qual(is) tipo(s) de relacionamento(s) deseja-se considerar na análise da curva de associação ao longo o tempo.

O SemMedDB permite conhecer a semântica de um relacionamento por meio do predicado que conecta dois termos (sujeito e objeto de uma sentença extraída do Medline). Existem 57 predicados que expressam relações no SemMedDB e são apresentados no Quadro 15.

Quadro 15: Predicados que expressam relações no SemMedDB.

ADMINISTERED_TO, AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, COEXISTS_WITH, COMPARED_WITH, COMPLICATES, CONVERTS_TO, DIAGNOSES, DISRUPTS, HIGHER_THAN, INHIBITS, INTERACTS_WITH, ISA, LOCATION_OF, LOWER_THAN, MANIFESTATION_OF, METHOD_OF, NEG_ADMINISTERED_TO, NEG_AFFECTS, NEG_ASSOCIATED_WITH, NEG_AUGMENTS, NEG_CAUSES, NEG_COEXISTS_WITH, NEG_COMPLICATES, NEG_CONVERTS_TO, NEG_DIAGNOSES, NEG_DISRUPTS, NEG_HIGHER_THAN, NEG_INHIBITS, NEG_INTERACTS_WITH, NEG_LOCATION_OF, NEG_LOWER_THAN, NEG_MANIFESTATION_OF, NEG_METHOD_OF, NEG_OCCURS_IN, NEG_PART_OF, NEG_PRECEDES, NEG_PREDISPOSES, NEG_PREVENTS, NEG_PROCESS_OF, NEG_PRODUCES, NEG_STIMULATES, NEG_TREATS, NEG_USES, OCCURS_IN, PART_OF, PRECEDES, PREDISPOSES, PREVENTS, PROCESS_OF, PRODUCES, SAME_AS, STIMULATES, TREATS, USES,

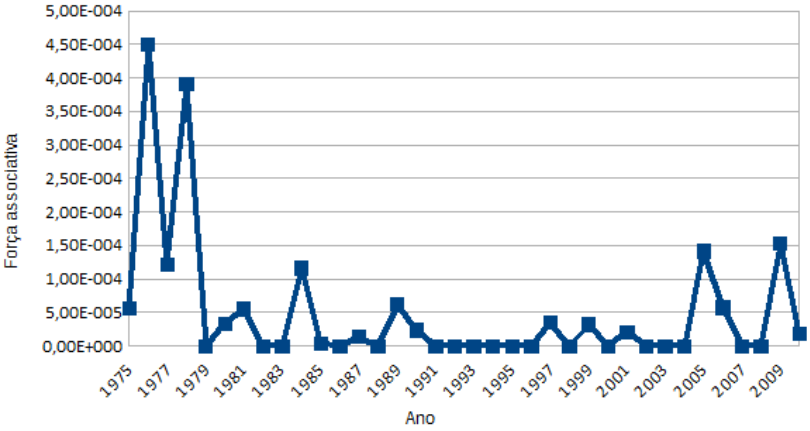
Fonte: Autor.

Dessa forma, foi realizado um experimento no qual os termos do Experimento 1 (associação entre os termos “Raynaud” e “Eicosapentaenoic”, vide Seção 5.2.1) se conectam apenas por um único predicado.

A escolha pelo predicado foi em função do maior número de ocorrência de tal predicado entre os 50 termos envolvidos no Experimento 1. O *ranking* é liderado pelo predicado COEXISTS_WITH seguido do predicado AFFECTS. Dessa forma, ambos foram escolhidos para verificar o comportamento das associações quando apenas um tipo de relacionamento é considerado.

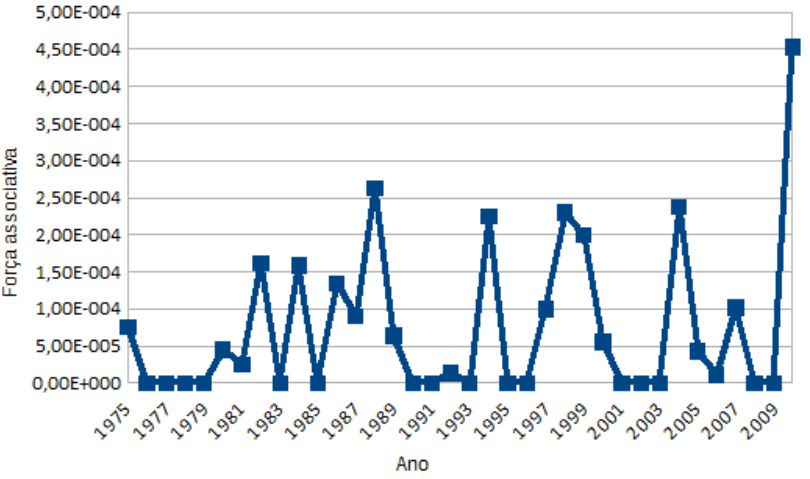
A Figura 20 e a Figura 21 apresentam a evolução da força associativa entre os termos "Raynaud" e "Eicosapentaenoic" entre 1975 a 2010 levando em conta apenas o relacionamento COEXISTS_WITH e AFFECTS, respectivamente.

Figura 20: Evolução da força associativa entre os termos "Raynaud" e "Eicosapentaenoic" entre 1975 a 2010 levando em conta apenas o relacionamento COEXISTS_WITH.



Fonte: Autor.

Figura 21: Evolução da força associativa entre os termos "Raynaud" e "Eicosapentaenoic" entre 1975 a 2010 levando em conta apenas o relacionamento AFFECTS.



Fonte: Autor.

Nestes experimentos, observa-se que a força associativa entre dois termos é muito baixa (em ambos os experimentos a maior foi de 0,00045), pois as matrizes são esparsas. A esparsidade das matrizes neste caso se deve ao fato de que, no SemMedDB, poucos pares de termos se conectam por um determinado relacionamento em um determinado ano.

Observa-se também que ambos os gráficos demonstram que a evolução de associação entre os dois termos em questão levando em conta apenas um tipo de relacionamento possui comportamento diferente quando comparada à evolução de associação que leva em conta todos os relacionamentos entre os dois termos. Também, a evolução de associação extraída a partir do relacionamento COEXISTS_WITH possui comportamento diferente da evolução de associação extraída a partir do relacionamento AFFECTS.

Isso demonstra que o modelo proposto permite aprimorar o tipo de análise a ser realizada e descobrir conhecimento a partir de diferentes pontos de vista.

5.2 CLASSIFICAÇÃO DE DOCUMENTOS

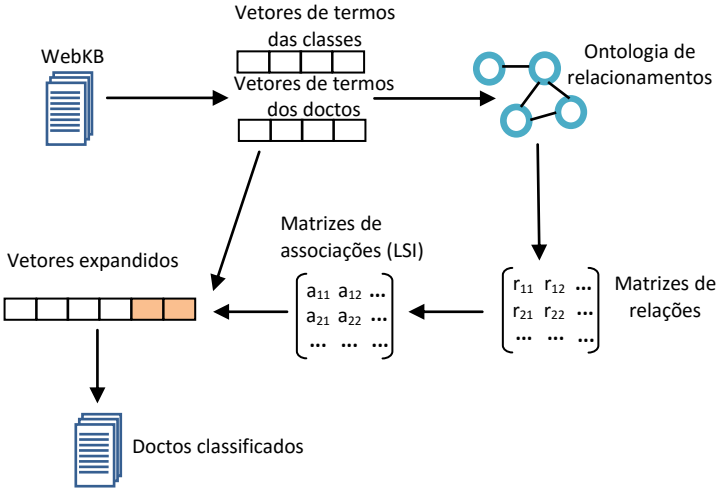
O modelo proposto foi aplicado na solução de um problema clássico de Aprendizagem de Máquina e Mineração de textos, a classificação, com o intuito de verificar a sua capacidade de generalização/adaptabilidade e como as associações podem contribuir para a classificação de textos.

O esquema do protótipo desenvolvido para a realização dos experimentos de classificação é mostrado na Figura 22.

A coleção de documentos utilizada para os testes foi o WebKB[7], um conjunto de páginas web de departamentos de Ciência da Computação de diversas universidades coletadas em 1997 pelo *CMU Text Learning Group* da *Carnegie Mellon School of Computer Science*. Este dataset contém 2803 arquivos categorizados em 4 classes: *Project* (336), *Student* (620), *Course* (750) e *Faculty* (1097).

[7] Disponível em: http://web.ist.utl.pt/acardoso/datasets/. Acesso em: 16/12/2015

Figura 22: Protótipo desenvolvido para os experimentos de classificação.



Fonte: Autor.

O processo foi realizado basicamente em 6 passos:

1 – Criação dos vetores de termos das classes e dos documentos

Para cada classe de documentos, foram identificados os termos de maior ocorrência em seus arquivos, organizados em um vetor. Isso foi realizado para reduzir a complexidade computacional da LSI (aplicada no passo 4), visto que os arquivos contém centenas de termos.

Os vetores de termos dos documentos contêm todos os termos presentes no documento.

2 – Instanciação da ontologia de relacionamentos

Após a identificação dos termos de cada classe, cria-se a ontologia de relacionamentos entre esses termos. É criada uma ontologia para cada classe.

Os pesos dos relacionamentos entre os termos da ontologia são obtidos a partir da coocorrência considerando todos os arquivos da classe. Os valores são normalizados dividindo-se todos pelo maior valor.

Nesse experimento os relacionamentos são considerados estáticos e, por isso, o fator tempo não é levado em conta, pois não

objetiva-se avaliar a evolução temporal dos relacionamentos e, além disso, a fonte de entrada WebKB não fornece essa informação.

3 – Matrizes de relações entre termos

As ontologias de relacionamentos são transformadas em matrizes *Termo X Termo.*

4 – Matrizes de associações entre termos

É aplicada a LSI sobre as matrizes de relações de cada classe de forma a identificar o grau de associação semântica entre os termos.

5 – Expansão dos vetores de termos de cada documento

Nesta etapa, é criada uma nova representação dos documentos a partir de termos associados. O vetor de termos de cada documento é expandido de acordo com a matriz de associação da sua respectiva classe. A partir disto são adicionados ao vetor original de termos do documento os termos associados com maior peso.

6 – Avaliação do Modelo

Na etapa de avaliação do modelo determina-se a aderência do mesmo em relação aos dados utilizados. Nessa fase utiliza-se a validação cruzada *Leave-One-Out* (DEVROYE; WAGNER, 1979). Cada documento da nova representação é comparado com todos os outros documentos, verificando sua similaridade através do cosseno. Se a classe do documento de maior similaridade for igual à classe do documento que está sendo comparado, contabiliza-se um acerto, caso contrário, um erro. O Quadro 16 apresenta os resultados dos testes realizados.

Quadro 16: Resultados da classificação de documentos utilizando o modelo proposto.

Vetor da classe / **Expansão dos vetores dos documentos**	**10 termos**	**20 termos**	**30 termos**
Expansão 0	65,60%	65,60%	65,60%
Expansão 2	66,96%	67,24%	66,96%
Expansão 4	68,49%	68,49%	68,78%
Expansão 7	68,96%	69,56%	69,63%
Expansão 10	69,10%	70,67%	70,85%

Fonte: Autor.

Foram realizados testes variando-se o tamanho do vetor de termos de cada classe (10, 20 e 30) e a expansão dos vetores dos documentos (sem expansão, expansão com 2, 4, 7 e 10 termos associados).

Os resultados mostram que na medida em que o tamanho do vetor de termos da classe e a expansão dos vetores dos documentos aumentam, a taxa de acertos também aumenta, o que demonstra a capacidade do modelo proposto ser aplicado como classificador de texto.

Esta pesquisa não objetiva comparar o modelo proposto com outros classificadores textuais, mas busca demonstrar sua capacidade de adaptabilidade por atender a mais de uma classe de problemas.

6 CONSIDERAÇÕES FINAIS

A extração de relacionamentos e associações entre elementos textuais constituem uma importante tarefa para a descoberta de conhecimento em bases textuais. Os detalhes semânticos de um relacionamento permitem a extração de associações relevantes para o levantamento de hipóteses potenciais para descoberta de conhecimento. A informação do tempo em que os relacionamentos ocorrem é igualmente importante para compreender a evolução desses relacionamentos e, consequentemente, do domínio sob investigação.

Uma revisão de literatura permitiu identificar uma lacuna nessa área: a inexistência de trabalhos que exploram, ao mesmo tempo, os detalhes semânticos e a dimensão temporal dos relacionamentos de forma a levantar associações relevantes que propiciem descoberta de conhecimento.

Assim, de forma a preencher a lacuna identificada na literatura, esta tese teve por objetivo propor um modelo de descoberta de conhecimento capaz de explorar os detalhes semânticos e a dimensão temporal dos relacionamentos para levantar associações entre elementos textuais (termos) de forma a gerar novas hipóteses e descobrir novos conhecimentos.

O modelo proposto conta com (i) uma ontologia de alto-nível criada para representar detalhes semânticos e temporais dos relacionamentos entre elementos textuais em um domínio do conhecimento e, (ii) a técnica de decomposição matricial *Latent Semantic Indexing* para determinar a força de associação entre elementos textuais que não se relacionam diretamente.

O processo de construção dessa tese foi operacionalizado pela metodologia *Design Science Research Metodology*, a qual se constitui em um processo rigoroso de projetar artefatos para resolver problemas, avaliar o que foi projetado ou o que está funcionando, e comunicar os resultados obtidos.

O método de avaliação empregado para avaliar o modelo foi o experimental. Dois tipos de experimentos foram realizados: um que trata da classificação de documentos e outro que trata da associação semântica e temporal entre termos.

Quanto à classificação de documentos, o intuito foi de verificar a capacidade de generalização/adaptabilidade do modelo e como as associações semânticas podem contribuir para a classificação de textos. A coleção de documentos utilizada para os testes foi o WebKB, um conjunto de páginas web de departamentos de Ciência da Computação

de diversas universidades coletadas em 1997 pelo *CMU Text Learning Group* da *Carnegie Mellon School of Computer Science*. Este dataset contém 2803 arquivos categorizados em 4 classes: *Project* (336), *Student* (620), *Course* (750) e *Faculty* (1097). Os resultados demonstraram o potencial do modelo proposto ser aplicado como classificador de texto, evidenciando sua capacidade de adaptabilidade por atender a mais de uma classe de problemas.

Quanto à associação, o intuito foi comprovar que o modelo é capaz de apresentar a evolução da força associativa entre dois termos ao longo do tempo, contribuindo para o levantamento de hipóteses e, consequentemente, para a descoberta de conhecimento. As bases de dados utilizadas nos experimentos de associação foram o Medline**,** uma base de dados bibliográfica e disponibiliza mais de 22 milhões de citações e resumos de publicações científicas das áreas da medicina, enfermagem, odontologia, medicina veterinária, biologia, bioquímica, evolução molecular, entre outros, e o SemMedDB, uma base de dados relacional contendo predicações semânticas (sujeito-predicado-objeto) extraídas dos títulos e resumos do Medline, bem como a informação temporal (ano) em que ocorre um relacionamento.

Os experimentos realizados para a avaliação da associação utilizam os contextos de duas descobertas seminais da DBL e demonstram que o aumento da força associativa entre termos ao longo do tempo pode indicar uma conexão direta desses termos no futuro, evidenciando o potencial do modelo para a descoberta de conhecimento.

Embora o modelo demonstre potencial para descoberta de conhecimento, os dados de entrada desempenham papel importante para essa tarefa. Conforme visto no experimento da Seção 5.1.3, para uma fonte de dados na qual as relações entre termos apresentam-se muito esparsas, não é possível ver uma evolução consistente das associações ao longo do tempo. Nesses casos, a curva apresenta comportamento irregular, permitindo apenas analisar pontos isoladamente no tempo, nos quais ocorrem picos.

Assim, dentre as principais contribuições deste trabalho, aponta-se um modelo para descoberta de conhecimento genérico que leva em conta os detalhes semânticos e a dimensão temporal dos relacionamentos potencializando a qualidade da geração de hipóteses. O principal objetivo é permitir a descoberta do conhecimento em diversos contextos em que as informações se encontram.

Outra importante contribuição reside na ontologia que integra o modelo. Atualmente, a representação tradicional de um relacionamento qualquer em uma ontologia resume-se a apenas dois nós e uma aresta

rotulada. Neste sentido, este trabalho contribui com a Engenharia de Conhecimento disponibilizando um artefato que evolui a representação de um relacionamento por permitir incluir seus detalhes semânticos e temporais.

Trabalhos futuros:

O modelo para descoberta de conhecimento proposto neste trabalho tem como saída um grande número de associações com seus respectivos pesos. Em um passo posterior, essa saída poderá ser melhor explorada por meio da análise de agrupamentos e de redes a fim de indicar grupos de termos a serem investigados na literatura por estarem fortemente associados uns com os outros. Ou ainda, termos isolados que conectam agrupamentos podem merecer uma investigação, pois podem ser pontos de interconexão entre literaturas ou áreas distintas, como é o caso da Bissociação. Em ambos os casos, podem ser apontados caminhos futuros levando a descoberta de conhecimento.

O potencial de descoberta de conhecimento proporcionado pelo modelo proposto também poderá ser melhor explorado por meio de técnicas de aprendizagem de máquina capazes de analisar a grande quantidade de associações resultantes e indicar possíveis conexões futuras entre dois termos.

Neste trabalho, a dimensão tempo está representada por ano. Isto se deve ao fato de que na área da Descoberta Baseada em Literatura, tal granularidade demonstra-se adequada. Contudo, outras áreas do conhecimento podem requerer uma granularidade menor (mês, semana, dia, etc.) podendo a ontologia para representação de relacionamentos ser estendida para atender a este requisito. A lógica temporal também pode ser explorada para especificar propriedades de domínios de conhecimentos dinâmicos.

PUBLICAÇÕES

FREITAS JUNIOR, Vanderlei; WOSZEZENKI, Cristiane; ANDERLE, Daniel Fernando; SPERONI, Rafael; NAKAYAMA, Marina Keiko. A pesquisa científica e tecnológica. **Espacios,** v. 35, n. 9, p. 12, 2014.

CECI, F.; WOSZEZENKI, C. R.; GONCALVES, A. L. O Uso de Anotações Semânticas e Ontologias para a Classificação de Documentos. **International Journal of Knowledge Engineering and Management**, v. 3, n. 5, p. 1-14, 2014.

FREITAS JUNIOR, V.; ANDERLE, D. F.; WOSZEZENKI, C. R. Portais como ferramenta de gestão do conhecimento em governo eletrônico: uma avaliação dos portais dos institutos federais do Estado de Santa Catarina. In: BOING, H. et al. (Orgs.). *Cadernos de pesquisa em inovação*: as novas tecnologias e as tendências em inovação. Florianópolis: PPGEGC/UFSC, 2013. p. 151-183

WOSZEZENKI, C. R., FREITAS JUNIOR, V., NAKAYAMA, M. **Inclusão Digital e Social: Exercendo a Cidadania com o Auxílio das Tecnologias de Informação e Comunicação**. International Conference on Interactive Computer aided Blended Learning. **Anais...**, Florianópolis, SC, 2013. p.1 – 6

WOSZEZENKI, C. R., FREITAS JUNIOR, V., ROVER, A. J. Inclusão Digital e Social: Cidadania e Autopoiese na Sociedade da Informação. **International Journal of Knowledge Engineering and Management**, v. 2, p. 94-108, 2013.

WOSZEZENKI, C. R., BESEN, F., SANTOS, J. L., STEIL, A. Desaprendizagem Organizacional: Uma revisão bibliométria e analítica da literatura. **Perspectivas em Gestão & Conhecimento**, v. 3, p. 128-147, 2013.

WOSZEZENKI, C. R.; GONCALVES, A. L. Mineração de textos biomédicos: uma revisão bibliométrica. **Perspectivas em Ciência da Informação**, v. 18, p. 24-44, 2013.

WOSZEZENKI, C. R.; FREITAS JUNIOR, V.; ANDERLE, D. F.; STEIL, A.; DANDOLINI, G. A.; SOUZA, J. A. O Twitter® como objeto de investigação empírico-quantitativa: uma revisão bibliométrica. **Espacios (Caracas)** , v. 34, p. 6, 2013.

NAZÁRIO, D. C.; WOSZEZENKI, C. R.; SELL, D.; GAUTHIER, F. O.; DANTAS, M. A. R. **Semantic Portals as a support for Knowledge Management in organizations**. International Conference on Information Systems and Technology Management. **Anais**..., São Paulo, SP, 2013.

WOSZEZENKI, C. R.; GONCALVES, A. L. Descoberta Baseada em Literatura: Estado da arte. **International Journal of Knowledge Engineering and Management**, v.1, p. 91-103, 2012.

WOSZEZENKI, C. R.; FREITAS JUNIOR, V.; CONSONI, D. P.; ANDERLE, D. F.; NAKAYAMA, M. **A Gestão do Conhecimento nos Institutos Federais do Estado de Santa Catarina**. XIX Simpósio de Engenharia de Produção. **Anais...,** Bauru, SP, 2012.

WOSZEZENKI, C. R.; BESEN, F.; SANTOS, J. L.; STEIL, A. **Desaprendizagem Organizacional: Uma revisão bibliométrica**. XIX Simpósio de Engenharia de Produção. **Anais...**, Bauru, SP, 2012.

WOSZEZENKI, C. R.; BESEN, F., SANTOS, J. L.; STEIL, A. **Mapeamento das Pubicações Acadêmico-Científicas sobre Desaprendizagem Organizacional**. 23º Simpósio Brasileiro de Informática na Educação. **Anais...**, Rio de Janeiro, RJ, 2012.

REFERÊNCIAS

ABATE, F.; ACQUAVIVA, A.; FICARRA, E.; PIVA, R.; MACII, E. Gelsius: a literature-based workflow for determining quantitative associations between genes and biological processes. **IEEE/ACM transactions on computational biology and bioinformatics**, v. 10, n. 3, p. 619–31, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24091396>. .

AGRAWAL, R.; MANNILA, H.; SRIKANT, R.; TOIVONEN, H.; VERKAMO, A. I. Fast Discovery of Association Rules. **Advances in knowledge discovery and data mining1**. p.307–328, 1996. Cambridge, MA: MIT Press.

AHLERS, C. B.; HRISTOVSKI, D.; KILICOGLU, H.; THOMAS, C. Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms. **Symposium A Quarterly Journal In Modern Foreign Literatures**, p. 6–10, 2007.

ALEMAN-MEZA, B.; NAGARAJAN, M.; RAMAKRISHNAN, C.; et al. Semantic Analytics on Social Networks : Experiences in Addressing the Problem of Conflict of Interest Detection. 15th International Conference on World Wide Web. **Anais...** p.407–416, 2006.

ALONSO, O.; GERTZ, M.; BAEZA-YATES, R. Clustering and exploring search results using timeline constructions. 18th ACM conference on Information and knowledge Management. **Anais...** , 2009.

ANYANWU, K.; MADUKO, A. SemRank : Ranking Complex Relationship Search Results on the Semantic Web. 14th International Conference on World Wide Web. **Anais...** p.117–127, 2005.

BAEZA-YATES, R.; RIBEIRO-NETO, B. A. **Modern information retrieval: the concepts and technology behind search**. 2nd ed. New York: ACM Press, 2001.

BERRAZEGA, I.; FAIZ, R. Identifying temporal relations between main events in new articles. 2013 ACS International Conference on Computer Systems and Applications (AICCSA). **Anais...** p.1–4, 2013. IEEE. Disponível em:

<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6616467>. .

BERRY, M. W.; DUMAIS, S. T.; BRIEN, G. W. O. Using Linear Algebra for Intelligent Information Retrieval. **SIAM Review**, v. 37, n. 4, p. 573–595, 1994.

BERTHOLD, M. R. Towards Bisociative Knowledge Discovery. In: M. R. Berthold (Ed.); **Bisociative Knowledge Discovery**. p.1–10, 2012. Springer-Verlag.

BORDIN, A. S. **Framework Baseado em Conhecimento para Análise da Colaboração Científica**, 2015. Tese (Doutorado em Engenharia e Gestão do Conhecimento). Universidade Federal de Santa Catarina (UFSC).

BÖTTCHER, M.; HÖPPNER, F.; SPILIOPOUOU, M. On exploiting the power of time in data mining. **SIGKDD Explor. Newsl.**, v. 10, n. 2, p. 3–11, 2008.

BOVO, A. B. **Um Modelo de Descoberta de Conhecimento Inerente à Evolução Temporal dos Relacionamentos Entre Elementos Textuais**, 2011. Tese (Doutorado em Engenharia e Gestão do Conhecimento). Universidade Federal de Santa Catarina (UFSC).

BRACHMAN, R. J.; HILL, M. The Future of Knowledge Reresentation. The Eighth National Conference on Artificial Intelligence (AAAI90). **Anais...** p.1082–1092, 1990. Boston, Massachussets.

BRACHMAN, R. J.; LEVESQUE, H. J. **Knowledge Representation and Reasoning**. Morgan Kaufmann Publishers, 2004.

BRUZA, P. Towards Operational Abduction from a Cognitive Perspective. **Logic Journal of IGPL**, v. 14, n. 2, p. 161–177, 2006. Disponível em: <http://jigpal.oxfordjournals.org/cgi/doi/10.1093/jigpal/jzk012>. Acesso em: 3/4/2013.

BUNDSCHUS, M.; DEJORI, M.; STETTER, M.; TRESP, V.; KRIEGEL, H.-P. Extraction of semantic biomedical relations from text using conditional random fields. **BMC bioinformatics**, v. 9, p. 207, 2008. Disponível em:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2386138&tool=pmcentrez&rendertype=abstract>. Acesso em: 25/3/2013.

BUNGE, M. **Treatise on Basic Philosophy. Part II: Life Science, Social Science and Technology**. Boston: D. Reidel, 1985.

ÇAĞDAŞ, V.; STUBKJÆR, E. Design research for cadastral systems. **Computers, Environment and Urban Systems**, v. 35, p. 77–87, 2011.

CAMERON, D.; BODENREIDER, O.; YALAMANCHILI, H.; et al. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. **Journal of biomedical informatics**, v. 46, n. 2, p. 238–51, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23026233>. .

CAMERON, D.; KAVULURU, R.; RINDFLESCH, T. C.; et al. Context-driven automatic subgraph creation for literature-based discovery. **Journal of biomedical informatics**, v. 54, p. 141–57, 2015. Elsevier Inc. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/25661592>. Acesso em: 8/12/2015.

CHAFFIN, R.; HERRMANN, D. J.; WINSTON, M. An empirical taxonomy of part-whole relations: Effects of part-whole type on relation identification. **Language and Cognitive Processes**, v. 3, n. 1, p. 17–48, 1988.

CHEN, E. S.; HRIPCSAK, G.; XU, H.; MARKATOU, M.; FRIEDMAN, C. Automated Acquisition of Disease – Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. **Journal of the American Medical Informatics Association**, v. 15, n. 1, 2008.

CHEN, H.; MARTIN, B.; DAIMON, C. M.; MAUDSLEY, S.; ZARAKET, F. A. Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. **Frontiers in Physiology**, v. 4, n. January, p. 1–6, 2013.

COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in bioinformatics**, v. 6, n. 1, p. 57–71, 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15826357>. .

COHEN, T.; SCHVANEVELDT, R. W. The trajectory of scientific discovery: concept co-occurrence and converging semantic distance. **Studies in Health Technology and Informatics**, v. 160, n. Pt 1, p. 661–5, 2010. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20841769>. .

CUPANI, A. La peculiaridad del conocimiento tecnológico. **ScientiaeStudia**, v. 4, n. 3, p. 353–71, 2006.

CUPANI, A. **A Filosofia da tecnologia: um convite**. UFSC: Florianópolis, 2011.

CVIJIKJ, I. P.; MICHAHELLES, F. Monitoring trends on Facebook. IEEE 9th International Conference on Dependable, Autonomic and Secure Computing. **Anais...** p.895–902, 2011. Sydney, Australia.

D'SOUZA, J.; NG, V. Classifying temporal relations in clinical data: A hybrid, knowledge-rich approach. **Journal of Biomedical Informatics**, v. 46, n. Supplement, p. S29–S39, 2013.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento Empresarial: como as organizações gerenciam o seu capital intelectual**. Rio de Janeiro: Campus, 1998.

DECKER, S.; MELNIK, S.; HARMELEN, F. VAN; et al. The Semantic Web : The Roles of XML and RDF. **IEEE Internet Computing**, v. 4, n. 5, p. 63–74, 2000.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by Latent Semantic Analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1998.

DEVROYE, L.; WAGNER, T. J. Distribution-free inequalities for the deleted and holdout error estimates. **Information Theory, IEEE Transactions on**, v. 25, n. 2, p. 202–207, 1979.

DUPONT, P.; CALLUT, J.; DOOMS, G.; MONETTE, J. N.; DEVILLE, Y. **Relevant subgraph extraction from random walks in a graph**. 2006.

EIJK, C. C. VAN DER; MULLIGEN, E. M. VAN; KORS, J. A.; MONS, B.; BERG, J. VAN DEN. Constructing an Associative Concept Space for Literature-Based Discovery. **Journal of the American Society for Information Science**, v. 55, n. 5, p. 436–444, 2004.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **Al Magazine**, v. 17, n. 3, p. 37–54, 1996.

FELDMAN, R.; DAGAN, I.; HIRSH, H. Mining text using keyword distributions. **Journal of Intelligent Information Systems**, v. 10, p. 281–300, 1998.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. New York: Cambridge University Press, 2006.

FREEMAN, L. C. Centrality in social networks conceptual clarification. **Social Networks**, v. 1, n. 3, p. 215–239, 1979.

FREITAS JUNIOR, V.; WOSZEZENKI, C.; ANDERLE, D. F.; SPERONI, R.; NAKAYAMA, M. K. A pesquisa científica e tecnológica. **Espacios**, v. 35, n. 9, p. 12, 2014.

GANDRA, P.; PRADHAN, M.; PALAKAL, M. J. Biomedical Association Mining and Validation. Proceedings of the International Symposium on Biocomputin. **Anais...** v. 1, p.1–8, 2010.

GANIZ, M. C.; POTTENGER, W. M.; JANNECK, C. D. Recent Advances in Literature Based Discovery. **Search**, 2005.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4th ed. São Paulo: Atlas, 2008.

GÓMEZ-PÉREZ, A. Ontological engineering: a state of the art. **British Computer Society**, v. 2, p. 33–43, 1999.

GONÇALVES, A.; ZHU, J.; SONG, D.; UREN, V.; PACHECO, R. LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval. **Lecture Notes in Computer Science**, v. 4016, p. 122–133, 2006.

GONG, L.; YANG, R.; SUN, X. BRES : EXTRACTING MULTICLASS BIOMEDICAL. **Biomedical Engineering: Applications, Basis and Communications**, v. 25, n. 1, p. 1–9, 2013.

GOODWIN, J. C.; COHEN, T.; COHENUTHTMCEDU, T.; RINDFLESCH, T. Discovery by scent : Discovery browsing system based on the Information Foraging Theory. IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). **Anais...** p.232–239, 2012.

GORDON, M. D.; DUMAIS, S. Using latent semantic indexing for literature based discovery. **Journal of the American Society for Information Science**, v. 49, n. 8, p. 674–685, 1998. Disponível em: <http://doi.wiley.com/10.1002/(SICI)1097-4571(199806)49:8<674::AID-ASI2>3.0.CO;2-T>. .

GREEN, R.; BEAN, C. A.; MYAENG, S. H. **The Semantic of Relationships: An Interdisciplinary Perspective**. Springer, 2002.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation. **Anais...** , 1993. Padova: Kluwer Academic Publishers.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, p. 199–220, 1993.

GUARINO, N. Formal ontology, conceptual analysis and knowledge representation. **International Journal of Human and Computer Studies**, v. 43, n. 5/6, p. 625–640, 1995.

GUARINO, N. Formal Ontology and Information Systems. FOIS'98. **Anais...** p.3–15, 1998. Trento, Italy: IOS Press.

GUO, W.; KRAINES, S. B. Mining Relationship Associations from Knowledge about Failures Using Ontology and Inference. In: Perner/Petra (Ed.); **Advances in Data Mining. Applications and Theoretical Aspects.** p.617–631, 2010. Springer-Verlag.

GUO, W.; KRAINES, S. B. Extracting Relationship Associations from Semantic Graphs in Life Sciences. **Communications in Computer and Information Science**, v. 128, p. 53–67, 2011.

HA-THUC, V.; MEJOVA, Y.; HARRIS, C.; SRINIVASAN, P. Event Intensity Tracking in Weblog Collections. 3rd International AAAI Conference on Weblogs and Social Media Data Challenge Workshop. **Anais...** , 2009. San Jose, California, USA.

HE, R.; QIN, B.; LIU, T.; LI, S. Cascaded Regression Analysis Based Temporal Multi-document Summarization. **Informatica - An International Journal of Computing and Informatics**, v. 34, n. 1, p. 119–124, 2010.

HEINZLE, R. **Um modelo de engenharia do conhecimento para sistemas de apoio a decisão com recursos para raciocínio abdutivo**, 2011. Universidade Federal de Santa Catarina.

HOMAYOUNI, R.; HEINRICH, K.; WEI, L.; BERRY, M. W. Gene clustering by latent semantic indexing of MEDLINE abstracts. **Bioinformatics (Oxford, England)**, v. 21, n. 1, p. 104–15, 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15308538>. Acesso em: 19/2/2014.

HRISTOVSKI, D.; FRIEDMAN, C.; RINDFLESCH, T. C.; PETERLIN, B. Exploiting semantic relations for literature-based discovery. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 349–53, 2006. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839258&tool=pmcentrez&rendertype=abstract>. .

HRISTOVSKI, D.; KASTRIN, A.; PETERLIN, B.; RINDFLESCH, T. C. Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation. **Lecture Notes in Bioinformatics**, p. 53–61, 2010.

HRISTOVSKI, D.; STARE, J.; PETERLIN, B.; DZEROSKI, S. Supporting discovery in medicine by association rule mining in Medline and UMLS. **Studies in health technology and informatics**, v. 84, n. Pt 2, p. 1344–8, 2001. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11604946>. .

HU, X. A Semantic-based Approach for Mining Undiscovered Public Knowledge from Biomedical Literature. ,2005.

HU, X.; ZHANG, X.; YOO, I.; ZHANG, Y. A Semantic Approach for Mining Hidden Links from Complementary and Non-interactive Biomedical Literature. , p. 200–209, 2006.

HU, X.; ZHANG, X.; YOO, I.; ZHOU, X.; XU, X. Mining Hidden Connections among Biomedical Concepts from Disjoint Biomedical Literature Sets through Semantic-Based Association Rule. **International Journal of Intelligent Systems**, v. 25, n. 2, p. 207–223, 2010.

HUNTER, L.; LU, Z.; FIRBY, J.; et al. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. **BMC bioinformatics**, v. 9, p. 78, 2008. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2275248&tool=pmcentrez&rendertype=abstract>. Acesso em: 14/3/2013.

IOANNOU, Z.-M.; MAKRIS, C.; PATRINOS, G. P.; TZIMAS, G. A set of novel mining tools for efficient biological. **Artificial Intelligence Review**, p. 1–18, 2013.

ITTIPANUVAT, V.; FUJITA, K.; KAJIKAWA, Y.; MORI, J.; SAKATA, I. Finding Linkage between Technology and Social Issues : A Literature Based Discovery Approach. PICMET ’12: Technology Management for Emerging Technologies. **Anais...** p.2310–2321, 2012.

KANG, B.-C.; KIM, H.-Y.; SHIN, G.-H.; et al. Semantic data integration to biological relationship among chemicals, diseases, and differential expressed genes. **BioChip Journal**, v. 5, n. 1, p. 63–71, 2011. Disponível em: <http://www.springerlink.com/index/10.1007/s13206-011-5110-7>. Acesso em: 3/4/2013.

KHAIRKAR, A. D.; KSHIRSAGAR, D. D.; KUMAR, S. Ontology for Detection of Web Attacks. Proceedings of the 2013 International Conference on Communication Systems and Network Technologies. **Anais...** p.612–615, 2013.

KHY, S.; ISHIKAWA, Y.; KITAGAWA, H. A Novelty-based Clustering Method for On-line Documents. **World Wide Web**, v. 11, n. 1, p. 1–37, 2008.

KILICOGLU, H.; SHIN, D.; FISZMAN, M.; ROSEMBLAT, G.; RINDFLESCH, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. **Bioinformatics (Oxford, England)**, v. 28, n. 23, p. 3158–60, 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23044550>. .

KIM, H.; PARK, H.; DRAKE, B. L. Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. **BMC bioinformatics**, v. 8 Suppl 9, p. S6, 2007. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2217664&tool=pmcentrez&rendertype=abstract>. Acesso em: 30/1/2014.

KIM, M. Concept map engineering: methods and tools based on the semantic relation approach. **Educational Technology Research and Development**, v. 61, n. 6, p. 951–978, 2013. Disponível em: <http://link.springer.com/10.1007/s11423-013-9316-3>. Acesso em: 3/12/2013.

KNELLER, G. F. **A Ciência como Atividade Humana**. Rio de Janeiro: Zahar, 1980.

KONTOSTATHIS, A.; POTTENGER, W. M.; PH, D. Detecting Patterns in the LSI Term-Term Matrix. IEEE International Conference on Data Mining (ICDM 02). **Anais...** , 2002. Maeybaski, Japan.

KURGAN, L. A.; MUSILEK, P. A survey of Knowledge Discovery and Data Mining process models. **The Knowledge Engineering Review**, v. 21, n. 01, p. 1, 2006. Disponível em: <http://www.journals.cambridge.org/abstract_S0269888906000737>. .

LACERDA, D. P.; DRESCH, A.; PROENÇA, A.; ANTUNES JÚNIOR, J. A. V. Design Science Research: método de pesquisa para a engenharia de produção. **Gest. Prod.**, v. 20, n. 4, p. 741–761, 2013.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. **Discourse Processes**, v. 25, n. 2-3, p. 259–284, 1998.

Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01638539809545028>. .

LANDIS, T. Y.; HEI-IMANN, D. J.; CHAFFIN, R. Development differences in the comprehension of semantic relations. **Zeitschrift für Psychologie**, v. 195, n. 2, p. 129–139, 1987.

LENT, B.; AGRAWAL, R.; SRIKANT, R. Discovering Trends in Text Databases. The Third International Conference on Knowledge Discovery and Data Mining. **Anais...** p.227–230, 1997. Newport Beach, California: AAAI Press.

LEROY, G.; CHEN, H. Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. **Journal of the American Society for Information Science and Technology**, v. 56, n. 5, p. 457–468, 2005. Disponível em: <http://doi.wiley.com/10.1002/asi.20135>. Acesso em: 3/4/2013.

LEVESQUE, H. J. KNOWLEDGE REPRESENTATION AND REASONING. **Annual Review of Computer Science**, v. 1, p. 255–287, 1986.

LI, G.; ZHANG, X.; YOO, I.; ZHOU, X. A Text Mining Method for Discovering Hidden Links. **Kobe Journal of Medical Sciences**, v. 55, n. 3, p. E53–E56, 2009.

LING, X.; WELD, D. S. Temporal Information Extraction. Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010). **Anais...** , 2010. Atlanta.

LIU, Y.; BILL, R.; FISZMAN, M.; et al. Using SemRep to Label Semantic Relations Extracted from Clinical Text. **AMIA Annual Symposium Proceedings**, p. 1–9, 2012.

LOGLISCI, C. Time-based discovery in biomedical literature: mining temporal links. **International Journal of Data Analysis Techniques and Strategies**, v. 5, n. 2, p. 148–174, 2011.

LOPES, L. F. **Um modelo de engenharia do conhecimento baseado em ontologia e cálculo probabilístico para o apoio ao diagnóstico**, 2011. Universidade Federal de Santa Catarina.

LUGER, G. F. **Inteligência Artificial: Estruturas e estratégias para a solução de problemas complexos**. Bookman, 2004.

MAGULURI, L. P.; KRISHNA, M. V.; SRIDHAR, P. S. S. A Novel Approach for Discovering Relevant Semantic Associations on Social Web Mining. Conference on IT in Business, Industry and Government (CSIBIG). **Anais...** p.1–7, 2014.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision Support Systems**, v. 15, n. 4, p. 251–266, 1995.

MCCALLA, G.; CERCONE, N. Guest Editors' Introduction: Approaches to Knowledge Representation. **Computer**, v. 16, n. 10, p. 12–18, 1983.

MCDONALD, D. M.; CHEN, H.; SU, H.; MARSHALL, B. B. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. **Bioinformatics (Oxford, England)**, v. 20, n. 18, p. 3370–8, 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15256411>. Acesso em: 15/3/2013.

MEI, Q.; ZHAI, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. **Anais...** , 2005. Chicago, Illinois, USA.

MENGLE, S. S. R.; GOHARIAN, N. Mining temporal relationships among categories. The 2010 ACM Symposium on Applied Computing. **Anais...** p.1107–1108, 2010. Sierre, Switzerland.

MINSKY, M. A framework for representing knowledge. In: P. Winston (Ed.); **The Psychology of Computer Vision**. p.211–277, 1975. New York: McGraw-Hill.

MONTES-Y-GÓMEZ, M.; GELBUKH, A.; LÓPEZ-LÓPEZ, A. Mining the News: Trends, Associations, and Deviations. **Computación y Sistemas**, v. 5, n. 1, p. 14–24, 2001.

MOSCHOPOULOS, T.; IOSIF, E.; DEMETROPOULOU, L.; POTAMIANOS, A.; NARAYANAN, S. Toward the Automatic Extraction of Policy Networks Using Web Links and Documents. **IEEE**

TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, v. 25, n. 10, p. 2404–2417, 2013.

MYLOPOULOS, J. An Overview of Knowledge Representation. Proceedings of the 1980 workshop on Data abstraction, databases and conceptual modeling. **Anais...** p.5–12, 1980. New York, NY: ACM.

OH, S.; YEOM, H. Y. A social network extraction based on relation analysis. **Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication - ICUIMC '12**, p. 1, 2012. New York, New York, USA: ACM Press. Disponível em: <http://dl.acm.org/citation.cfm?doid=2184751.2184805>. .

PACHECO, R. C. DOS S. DADOS E GOVERNO ABERTOS NA SOCIEDADE DO CONHECIMENTO. Disponível em: <http://www.inf.ufsc.br/~tite/LODBrasil/Abertura/DadosEGovernoAbertoNaSocConh.pdf>. Acesso em: 3/3/2016.

PACHECO, R. C. DOS S.; TOSTA, K. C. B. T.; FREIRE, P. DE S. Interdisciplinaridade vista como um processo complexo de construção de conhecimento: uma análise do Programa de Pós-Graduação EGC/UFSC. **RBPG-Revista Brasileira de Pós-Graduação**, v. 7, n. 12, p. 136–159, 2010.

PAPADIMITRIOU, C. H.; TAMAKI, H.; RAGHAVAN, P.; VEMPALA, S. Latent Semantic Indexing : A Probabilistic Analysis. Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. **Anais...** p.159–168, 1998. New York: ACM Press.

PATEL, C. O.; CIMINO, J. J. Using semantic and structural properties of the Unified Medical Language System to discover potential terminological relationships. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 3, p. 346–53, 2009. J Am Med Inform Assoc. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3202330&tool=pmcentrez&rendertype=abstract>. Acesso em: 28/3/2013.

PAUL, R.; GROZA, T.; HUNTER, J.; ZANKL, A. Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain. **Journal of biomedical semantics**, v. 5, n. 1, p. 8, 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24499729>. .

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45–78, 2007.

PREISS, J.; STEVENSON, M.; MCCLURE, M. H.; COURT, R. **Towards Semantic Literature Based Discovery**. 2012.

RAGHAVAN, P.; FOSLER-LUSSIER, E.; LAI, A. M. Learning to temporally order medical events in clinical text. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012. **Anais...** p.70–74, 2012.

RAUTENBERG, S. **Modelo de conhecimento para mapeamento de instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias**, 2009. Universidade Federal de Santa Catarina.

RINDFLESCH, T. C.; FISZMAN, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. **Journal of biomedical informatics**, v. 36, n. 6, p. 462–77, 2003. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/14759819>. Acesso em: 26/3/2013.

RINDFLESCH, T. C.; TANABE, L.; WEINSTEIN, J. N.; HUNTER, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. **Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**, p. 517–28, 2000. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2709525&tool=pmcentrez&rendertype=abstract>. .

ROY, S.; HEINRICH, K.; PHAN, V.; BERRY, M. W.; HOMAYOUNI, R. Latent Semantic Indexing of PubMed abstracts for identification of transcription factor candidates from microarray derived gene sets. **BMC bioinformatics**, v. 12, n. Suppl 10, p. S19, 2011. BioMed Central Ltd. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3236841&tool=pmcentrez&rendertype=abstract>. Acesso em: 19/2/2014.

SAVOVA, G.; BETHARD, S.; STYLER, W.; et al. Towards temporal relation discovery from the clinical narrative. AMIA 2009 Annual Symposium. **Anais...** p.568–572, 2009.

SCHREIBER, G.; AKKERMANS, H.; ANJEWIERDEN, A.; et al. **Knowledge Engineering and Management: The CommonKADS Methodology**. Cambridge: The MIT Press, 2002.

SHARMA, A.; SWAMINATHAN, R.; YANG, H. A Verb-centric Approach for Relationship Extraction in Biomedical Text. IEEE Fourth International Conference on Semantic Computing (ICSC). **Anais...** p.377–385, 2010. Pittsburgh, PA.

SHETH, A.; ALEMAN-MEZA, B.; ARPINAR, I. B.; HALASCHEK, C.; RAMAKRISHNAN, C. Semantic Association Identification and Knowledge Discovery for National Security Applications. **Journal of Database Management**, v. 16, n. 1, p. 33–53, 2005. Springer-Verlag.

SHORTLIFFE, E. H. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. Proceedings of the 1974 annual ACM conference. **Anais...** p.739–739, 1974.

SIMON, H. **The Sciences of the Artificial**. 3rd ed. Cambridge/Massachussets: MIT Press, 1996.

SMALHEISER, N. R.; SWANSON, D. R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. **Computer methods and programs in biomedicine**, v. 57, n. 3, p. 149–53, 1998. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9822851>. .

SOWA, J. F. **Knowledge Representation: Logical, Philosophical, and Computational Foundations**. 1st ed. Pacific Grove: Brooks Cole Publishing Co., 2000.

SRINIVASAN, P. Text Mining : Generating Hypotheses from MEDLINE. **Journal of the American Society for Information Science and Technology**, v. 55, n. 5, p. 396–413, 2004. Disponível em: <http://doi.wiley.com/10.1002/asi.10389>. .

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1-2, p. 161–197, 1998.

SUBASIC, I.; BERENDT, B. Web Mining for Understanding Stories through Graph Visualisation. The 8th IEEE International Conference on Data Mining. **Anais...** p.570–579, 2008. Pisa, Italy.

SUCHITRA, A.; SUDHA, R. Extraction of Semantic Biomedical Relations from Medline Abstracts using Machine Learning Approach. National Conference on Advances in Computer Science and Applications (NCACSA 2012). **Anais...** p.1–4, 2012.

SULLIVAN, D. **Document Warehousing and Text Mining**. New York: John Wiley & Sons, Inc., 2001.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. **Journal of the American Medical Informatics Association**, v. 20, n. 5, p. 806–813, 2013.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Temporal reasoning over clinical text: the state of the art. **Journal of the American Medical Informatics Association**, v. 20, n. 5, p. 814–9, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23676245>. Acesso em: 14/1/2014.

SWANSON, D. R. Fish Oil, Raynaud's Syndrome, amd undiscovered public knowledge. **Perspectives in Biology and Medicine**, p. 7–18, 1986.

SWANSON, D. R. Migraine and Magnesium: eleven neglected connections. **Perspectives in Biology and Medicine**, v. 31, n. 4, p. 526–557, 1988.

TAMBA-MECZ, I. **A semântica**. São Paulo, 2006.

TANG, B.; WU, Y.; JIANG, M.; et al. A hybrid system for temporal information extraction from clinical text. **Journal of the American Medical Informatics Association**, v. 20, n. 5, p. 828–835, 2013.

TSAI, R. T.-H.; CHOU, W.-C.; SU, Y.-S.; et al. BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model

with automatically generated template features. **BMC bioinformatics**, v. 8, p. 325, 2007. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2072962&tool=pmcentrez&rendertype=abstract>. Acesso em: 3/4/2013.

TSURUOKA, Y.; MIWA, M.; HAMAMOTO, K.; TSUJII, J.; ANANIADOU, S. Discovering and visualizing indirect associations between biomedical concepts. **Bioinformatics**, v. 27, n. 13, p. i111–9, 2011. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3117364&tool=pmcentrez&rendertype=abstract>. Acesso em: 5/10/2013.

VARGAS, M. **Metodologia da pesquisa tecnológica**. Rio de Janeiro, 1985.

VASSEV, E.; HINCHEY, M. Knowledge Representation and Reasoning for Intelligent Software Systems. **Computer**, v. 44, n. 8, p. 96–99, 2011.

VENZIN, M.; KROGH, G. VON; ROOS, J. Future Research into Knowledge Management. In: G. von Krogh; J. Roos; D. Kleine (Eds.); **Knowing in Firms: Understanding, Managing and Measuring Knowledge**, 1998.

VIDAL, M.-E.; RASHID, L.; IBÁÑEZ, L.; et al. A Ranking-Based Approach to Discover Semantic Associations Between Linked Data. 4th Asian conference on Intelligent Information and Database Systems. **Anais...** p.152–162, 2010. Springer Berlin / Heidelberg.

VOGELSTEIN, E.; LANE, D.; LEVINE, A. J. Surfing the p53 network. **Nature**, v. 408, n. 6810, p. 307–310, 2000.

WAZLAWICK, R. S. Uma Reflexão sobre a Pesquisa em Ciência da Computação à Luz da Classificação das Ciências e do Método Científico. **Revista de Sistemas de Informação da FSMA**, v. 6, p. 3–10, 2010.

WEEBER, M. Advances in literature-based discovery. **2007-10-22]. http://nl. ijs. si/et/talks/tsujiilab/saso/**, 2003. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4217&rep=rep1&type=pdf>. Acesso em: 10/4/2012.

WEEBER, M.; KLEIN, H.; JONG-VAN DEN BERG, L. T. W. DE; VOS, R. Using concepts in literature-based discovery: Simulating Swanson's

Raynaud-fish oil and migraine-magnesium discoveries. **Journal of the American Society for Information Science and Technology**, v. 52, n. 7, p. 548–557, 2001. Disponível em: <http://doi.wiley.com/10.1002/asi.1104>. .

WEISSENBORN, D.; SCHROEDER, M.; TSATSARONIS, G. Discovering relations between indirectly connected biomedical concepts. **Journal of biomedical semantics**, v. 6, p. 28, 2015. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4492092&tool=pmcentrez&rendertype=abstract>. Acesso em: 3/12/2015.

WILKOWSKI, B.; FISZMAN, M.; PH, D.; et al. Graph-Based Methods for Discovery Browsing with Semantic Predications. AMIA Symposium. **Anais...** p.1514–1523, 2011.

WINSTON, M. E.; CHAFFIN, R.; HERRMANN, D. A taxonomy of part-whole relations. **Cognitive Science**, v. 11, p. 417–444, 1987.

WOSZEZENKI, C. R.; GONÇALVES, A. L. Mineração de Textos Biomédicos: Uma revisão bibliométrica. **Perspectivas em Ciência da Informação**, v. 18, n. 3, p. 24–44, 2013.

WREN, J. D. Using fuzzy set theory and scale-free network properties to relate MEDLINE terms. **Soft Computing**, v. 10, n. 4, p. 374–381, 2006. Disponível em: <http://link.springer.com/10.1007/s00500-005-0497-5>. Acesso em: 5/2/2014.

WU, Z.; ZHOU, X.; LIU, B.; CHEN, J. Text Mining for Finding Functional Community of Related Genes Using TCM Knowledge. **Lecture Notes in Computer Science**, v. 3202, p. 459–470, 2004.

XU, Z.; LUO, X.; ZHANG, S.; et al. Mining temporal explicit and implicit semantic relations between entities using web search engines. **Future Generation Computer Systems**, 2013. Elsevier B.V. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167739X13002069>. Acesso em: 3/12/2013.

YANG, H.; SWAMINATHAN, R.; SHARMA, A.; KETKAR, V.; SILVA, D. Mining Biomedical Text Towards Building a Quantitative Food-disease-gene Network. , p. 205–225, 2011.

ZHU, S.; OKUNO, Y.; TSUJIMOTO, G.; MAMITSUKA, H. Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and medline. **Cancer informatics**, v. 2, p. 361–71, 2006. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2675505&tool=pmcentrez&rendertype=abstract>. .