**UNIVERSIDADE FEDERAL DE SANTA CATARINA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO**

Alessandro Botelho Bovo

# UM MODELO DE DESCOBERTA DE CONHECIMENTO INERENTE À EVOLUÇÃO TEMPORAL DOS RELACIONAMENTOS ENTRE ELEMENTOS TEXTUAIS

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

Florianópolis

2011

Catalogação na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

B783m Bovo, Alessandro Botelho
Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais [tese] / Alessandro Botelho Bovo ; orientador, Vinícius Medina Kern. – Florianópolis, SC, 2011.
155 p.: il., tabs.

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. Sistemas de recuperação da informação - Avaliação. 3. Redes de informação - Pesquisa - Fontes de informação - Estudo de casos. I. Kern, Vinicius Medina. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. III. Título.

CDU 659.2

Alessandro Botelho Bovo

**UM MODELO DE DESCOBERTA DE CONHECIMENTO INERENTE À EVOLUÇÃO TEMPORAL DAS RELAÇÕES ENTRE ELEMENTOS TEXTUAIS**

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

Florianópolis, 2 de fevereiro de 2011

_____
Prof. Paulo Maurício Selig, Dr.
Coordenador do Curso

**Banca examinadora:**

| | |
|---|---|
| _____<br>Vinícius Medina Kern, Dr.<br>Orientador<br>UFSC | _____<br>Aran Bey Tcholakian Morales, Dr.<br>Membro<br>UFSC |
| _____<br>Ricardo Pietrobon, Dr.<br>Examinador externo<br>Duke University | _____<br>Cláudio Chauke Nehme, Dr.<br>Examinador externo<br>UCB |
| _____<br>José Leomar Todesco, Dr.<br>Membro<br>UFSC | _____<br>Roberto Carlos dos Santos Pacheco, Dr.<br>Membro<br>UFSC |

## AGRADECIMENTOS

A Deus, por me acompanhar e iluminar o meu caminho.

Ao meu orientador, Vinícius Medina Kern, e aos professores Roberto Carlos dos Santos Pacheco e José Leomar Todesco, do EGC, e ao professor Ricardo Pietrobon, da *Duke University*, pelas contribuíções ao desenvolvimento da pesquisa.

Ao meu coorientador, Alexandre Leopoldo Gonçalves, que teve participação fundamental na elaboração desta tese.

Aos amigos do Instituto Stela, pelos vários anos de trabalho em conjunto.

Ao Instituto Stela, pelo apoio à realização da pesquisa.

Ao professor Barend Mons, do *Leiden University Medical Center* (LUMC) e do *University Medical Center of Rotterdam* (ErasmusMC), pela orientação durante o período sanduíche no LUMC.

Ao colega Herman van Haagen, do LUMC, pelo apoio dado durante minha estada na Holanda.

Aos pesquisadores Peter-Bram 't Hoen, Rob Jelier e Christine Chichester, do LUMC, e Martijn Schuemie, Erik van Mulligen e Jan Kors, do ErasmusMC, pelas contribuíções científicas à minha pesquisa.

Aos meus pais, Getúlio e Leidí, e aos meus irmãos, Fábio e Eduardo, que, mesmo estando longe, foram fundamentais para que eu conseguisse chegar até aqui.

Aos meus sogros, Carlos e Maria, cujo apoio foi fundamental para término desta tese.

E em especial à minha esposa, Alessandra, e à minha filha, Júlia.

## RESUMO

Há algum tempo tem sido observado e discutido o aumento expressivo na quantidade de informação produzida e publicada pelo mundo. Se por um lado essa situação propicia muitas oportunidades de uso dessa informação para a tomada de decisão, por outro, lança muitos desafios em como armazenar, recuperar e transformar essa informação em conhecimento. Umas das formas de descoberta de conhecimento que tem atraído atenção de pesquisadores é a análise dos relacionamentos presentes nas informações disponíveis. Não obstante, devido à grande velocidade de criação de novos conteúdos a dimensão *tempo* torna-se uma propriedade intrínseca e relevante presente nestas fontes de informação. Assim, o objetivo é desenvolver um modelo para descoberta de conhecimento a partir de informações não estruturadas analisando a evolução dos relacionamentos entre os elementos textuais ao longo do tempo. O modelo proposto é dividido por fases, assim como os modelos tradicionais de descoberta de conhecimento. As fases deste modelo são: configuração dos temas de análise, identificação das ocorrências dos conceitos, correlação e correlação temporal, associação e associação temporal, criação do repositório de temas de análise, e tarefas intensivas em conhecimento, com ênfase nos relacionamentos diretos e indiretos entre os conceitos do domínio. A demonstração de viabilidade é realizada por meio de um protótipo baseado no modelo proposto e sua aplicação em um estudo de caso. É realizada também uma análise comparativa do modelo proposto com outros modelos de descoberta de conhecimento em textos.



**Palavras-chave:** Descoberta de Conhecimento em Textos, *Temporal Knowledge Discovery in Texts*, *Temporal Text Mining*, Correlação de Elementos Textuais, Associação de Elementos Textuais, Análise Temporal de Informações Textuais.

## ABSTRACT

It has been observed and discussed the significant increase in the amount of information produced and published worldwide. On the one hand, this situation provides many opportunities to use this information for decision making, and on the other hand it throws many challenges on how to store, retrieve and transform that information into knowledge. One way of discovering knowledge that has attracted attention of researcher is the analysis of concept relationships present in the information. Nevertheless, due to the fast creation of new content the *time* dimension has become an intrinsic and relevant property present in these information sources. Thus, the aim is to develop a model for knowledge discovery from unstructured information by analyzing the evolution of relationships between textual concepts over time. The proposed model is divided in steps, as well as the traditional knowledge discovery models. The model steps are the following: setting the analysis themes, identifying occurrences of concepts, correlation and temporal correlation, association and temporal association, the creation of the themes analysis repository, and knowledge-intensive tasks with emphasis on direct and indirect relationships between domain concepts. A feasibility demonstration is performed by a prototype based on the proposed model and its application in a case study. It also performed a comparative analysis of the proposed model with other knowledge discovery in texts models.



**Keywords:** Knowledge Discovery in Texts, Temporal Knowledge Discovery in Texts, Temporal Text Mining, Correlation of Textual Concepts, Association of Textual Concepts, Temporal Analysis of Textual Information.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE SIGLAS

| | |
|---|---|
| AT | Análise de Tendências |
| DBL | Descoberta Baseada em Literatura |
| DRT | Detecção e Rastreamento de Tópicos |
| DTE | Detecção de Tendências Emergentes |
| AT | Análise de Tendências |
| DW | *Data Warehouse* |
| EC | Engenharia do Conhecimento |
| EGC | Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento |
| EI | Extração de Informação |
| GC | Gestão do Conhecimento |
| IA | Inteligência Artificial |
| IM | Informação Mútua |
| ISL | Indexação de Semântica Latente |
| KDD | *Knowledge Discovery in Databases* |
| KDT | *Knowledge Discovery in Texts* |
| LRD | *Latent Relation Discovery* |
| MD | Mineração de Dados |
| MEV | Modelo Espaço Vetorial |
| MT | Mineração de Textos |
| NER | *Named Entity Recognition* |
| OWL | *Web Ontology Language* |
| PLN | Processamento de Linguagem Natural |
| RI | Recuperação de Informação |
| SBC | Sistema Baseado em Conhecimento |
| TF-IDF | *Term Frequency - Inverted Document Frequency* |
| TKDT | *Temporal Knowledge Discovery in Texts* |
| TTM | *Temporal Text Mining* |
| UFSC | Universidade Federal de Santa Catarina |

**SUMÁRIO**

# 1 INTRODUÇÃO

Há algum tempo tem sido observado e discutido o aumento expressivo na quantidade de informação produzida e publicada pelo mundo. Segundo alguns autores (GREENGRASS, 2000; KOBAYASHI; TAKEDA, 2000; LYMAN, 2000; 2003; HIMMA, 2007), esse aumento tem ocorrido em escala exponencial. Tal situação se deve principalmente aos avanços nas tecnologias da informação e comunicação que, além de promover o aumento na quantidade, está fazendo com que essas informações se tornem cada vez mais acessíveis aos indivíduos e organizações. A pesquisa realizada por Lyman (2000) apontou que a quantidade de conteúdos disponíveis na Internet duplicava anualmente, e estimou em mais de dois bilhões o número páginas disponíveis na Internet no início do ano 2000. Smyth et al. (2004), a partir dos estudos de Lyman (2000; 2003), forneceram uma estimativa da existência de 10 bilhões de documentos. Shaw (2005) estimou em aproximadamente treze bilhões o número de páginas em 2005.

Mesmo antes da invenção da Internet, e mais especificamente da *World Wide Web*, já se notava o rápido aumento na quantidade de informações em áreas acadêmicas e não acadêmicas. Em 1987, Warren Thorngate observou que a quantidade de informações técnicas e científicas publicadas dobrava a cada período de 5 a 15 anos (HIMMA, 2007). As publicações científicas são responsáveis por uma parcela significativa da quantidade de informações produzidas atualmente. Por exemplo, o MEDLINE[1], que é um banco de dados bibliográfico com artigos científicos publicados nas áreas de ciências da biomédicas (medicina, farmácia, etc.), contém mais 18 milhões de registros[2].

Além de artigos científicos e a Web, há ainda vários outros tipos de informação textual em formato digital dentro das organizações: (a) os diversos tipos de relatórios técnicos, que podem conter muitas informações importantes sobre suas atividades, as quais podem ser úteis para se descobrir erros cometidos, soluções encontradas, quem fez o quê, etc.; (b) manuais disponíveis sobre procedimentos, softwares, etc.;

1 http://www.ncbi.nlm.nih.gov/pubmed/

2 http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update, em 04 de Outubro de 2010.

(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 (*search engines*), como o Google®, podem conter informações úteis sobre os interesses e necessidades dos seus colaboradores. Além desses exemplos, há ainda outros tipos de informações não estruturadas dentro das organizações: currículos, *e-books*, mensagens de comunicação instantânea, etc.

Se por um lado essa situação propicia muitas oportunidades de uso dessa informação para a tomada de decisão, por outro, lança muitos desafios em como armazenar, recuperar e transformar essa informação em *conhecimento*. Segundo Levy (2005; 2006), o problema de se lidar com muita informação é que se perde um tempo que poderia ser melhor empregado pensando, contemplando e raciocinando. A superação dos desafios de como obter conhecimento a partir desse excesso informações pode significar vantagem competitiva para as organizações.

Esses avanços nas tecnologias de comunicação e compartilhamento de informações, o aumento expressivo na quantidade e na importância destas informações, e a necessidade de transformar essas informações em conhecimento para as organizações, contribuíram para o surgimento da Gestão do Conhecimento (GC) (STUDER; DECKER et al., 2000; HOLSAPPLE, 2005). Nesse contexto, a Engenharia do Conhecimento (EC) é a área responsável por métodos e ferramentas que possibilitem o desenvolvimento de Sistemas Baseados em Conhecimento (SBC) para apoiar os diversos processos (criação, organização, formalização, compartilhamento, aplicação e refinamento) da GC (HENDRIKS, 1999; STUDER; DECKER et al., 2000; SCHREIBER; AKKERMANS et al., 2002; NISSEN, 2006).

No contexto da EC, áreas relativas à extração e recuperação da informação e descoberta de conhecimento desempenham um papel cada vez mais importante no desenvolvimento de SBCs. Como afirmam Hair et al. (1998), a área de descoberta de conhecimento se baseia na grande quantidade de informações disponíveis como também em questionamentos sobre essa informação. Assim, a análise de dados passa a ter um caráter mais exploratório, visando identificar ou explicitar conhecimento oculto em bases de dados. Essa tarefa é tradicionalmente de responsabilidade da área de Descoberta de Conhecimento em Bases de Dados (KDD – *Knowledge Discovery in Databases*) e de sua principal etapa, a Mineração de Dados (MD). Segundo Fayyad (1996),

KDD é um processo não trivial de identificação, a partir de dados, de padrões novos, válidos, potencialmente úteis e compreensíveis. Um dos seus principais passos é a MD, sendo esta responsável pela aplicação de algoritmos com o propósito de identificar padrões em uma base de dados (FAYYAD; PIATETSKY-SHAPIRO et al., 1996a). A MD pode ser entendida como "a exploração e a análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados, com o objetivo de descobrir padrões e regras significantes" (BERRY; LINOFF, 1997).

A área de descoberta de conhecimento em bases de dados e a mineração de dados lidam usualmente com dados estruturados. Contudo, como pode ser observado nos números apresentados anteriormente, a maior parte das informações atualmente disponíveis para as organizações são *não estruturadas*. Deste modo, devido à necessidade de se dar maior ênfase a dados não estruturados, houve a necessidade de se adaptar os métodos tradicionais de KDD e MD surgindo, assim, a Descoberta de Conhecimento em Textos (KDT – *Knowledge Discovery in Texts*) e a Mineração de Textos (MT). As abordagens de KDT e MT utilizam-se de métodos oriundos de áreas como Processamento de Linguagem Natural (PLN), Extração de Informação (EI), Recuperação de Informação (RI) e da Estatística.

Umas das formas de descoberta de conhecimento que tem atraído atenção de pesquisadores é a análise das interconexões presentes nas informações disponíveis. Tais interconexões podem representar, por exemplo, redes de relacionamentos (LIPNAK; STAMP, 1992; WEISZ; ROCO, 1996; NEWMAN, 2001; BARABÁSI, 2003; BALANCIERI; BOVO et al., 2005), comunidades virtuais de prática (RHEINGOLD, 1994; WENGER; MCDERMOTT et al., 2002; TERRA, 2003), ou até mesmo interações entre proteínas (STELZL; WORM et al., 2005; EOGHAN; LARS et al., 2008; VAN HAAGEN; 'T HOEN et al., 2009; BROWNE; ZHENG et al., 2010; NIU; OTASEK et al., 2010). Nesse sentido, uma das abordagens de KDT e MT consiste no desenvolvimento de modelos, métodos, técnicas e algoritmos para descoberta de conhecimento em bases de dados textuais a partir da análise dos relacionamentos entre elementos textuais (conceitos, termos, palavras, etc.) de um domínio. Como afirma Gonçalves (2006), essas abordagens promovem uma estrutura geral para revelar conhecimento oculto em coleções de documentos textuais e como esse conhecimento pode auxiliar no entendimento das relações estabelecidas intra e interorganização. Esses métodos podem ser classificados em dois

grupos: (a) métodos para descoberta de conhecimento baseados na *correlação* de elementos textuais; e (b) métodos para descoberta de conhecimento baseados na *associação* de elementos textuais.

A correlação mostra o relacionamento *direto* entre dois elementos textuais baseado em suas coocorrências nos documentos. Várias pesquisas em MT estão voltadas à análise de relacionamentos diretos entre conceitos em informações textuais (ZHU; GONÇALVES et al., 2005; ERHARDT; SCHNEIDER et al., 2006; GONÇALVES; ZHU et al., 2006; GONÇALVES; BEPPLER et al., 2006; ZHU; GONÇALVES et al., 2007; DAVIDOV; RAPPOPORT, 2008; GARTEN; ALTMAN, 2009; YAN; MATSUO et al., 2009; BUI; NUALLAIN et al., 2010; CHEN, 2010; MESQUITA; MERHAV et al., 2010; ROSE; ENGEL et al., 2010; SÁNCHEZ, 2010). Para se analisar tais relacionamentos, a forma mais simples consiste em contar o número de coocorrências entre os elementos textuais. Contudo, existem métodos mais sofisticados que permitem determinar o *peso* do relacionamento. Para o cálculo desse peso, utilizam-se alguns modelos com origem na estatística descritiva, tais como o teste *t*, *Chi-square* ($x^2$) e o *Z score* (MANNING; SCHÜTZE, 1999); os de origem na teoria da informação, tais como Informação Mútua (IM) (CHURCH; HANKS, 1990; CHURCH; GALE, 1991) e *Phisquared* ($\Phi^2$) (CHURCH; GALE, 1991); os modelos com base mais empírica como o algoritmo CORDER (ZHU; GONÇALVES et al., 2005) e o *Latent Relation Discovery* (LRD) (GONÇALVES; ZHU et al., 2006); e tem-se ainda o modelo Indexação de Semântica Latente (ISL), que tem por objetivo capturar a estrutura semântica de coleções de documentos através da correlação de termos e documentos (DEERWESTER; DUMAIS et al., 1990; DING, 2000).

A associação mostra o relacionamento *indireto* entre dois elementos textuais baseado nos contextos nos quais eles aparecem nos documentos. Para o seu cálculo, utilizam-se, numa primeira etapa, modelos baseados em coocorrência para, numa etapa posterior, tentar identificar relacionamentos indiretos entre aqueles elementos que não coocorrem, ou que coocorrem com uma frequência muito baixa, através do contexto de cada elemento. Os trabalhos relativos a estes métodos ─ que são genericamente chamados de Descoberta Baseada em Literatura (DBL) ─ são em sua grande maioria aplicados em informações textuais das ciências biomédicas (SWANSON, 1986; WEEBER, 2003; GANIZ; POTTENGER et al., 2006; VAN HAAGEN; 'T HOEN et al., 2009; BAKER; HEMMINGER, 2010; COHEN; SCHVANEVELDT et al.,

2010; GANDRA; PRADHAN et al., 2010; ZHOU; PENG et al., 2010). Todavia, segundo Weeber (2003), a DBL pode ser aplicada em textos de qualquer área de conhecimento. Nessa mesma linha, Ganiz et al. (2006) afirmam que, apesar de correntemente ser utilizada principalmente no domínio das ciências biomédicas, a DBL tem um amplo potencial de aplicação.

Não obstante, devido à grande velocidade de criação de novos conteúdos – como discutido anteriormente – a dimensão tempo torna-se uma propriedade intrínseca e relevante presente nestas informações (KHY; ISHIKAWA et al., 2008; SUBASIC; BERENDT, 2008; ABE; TSUMOTO, 2009; BERENDT; SUBASIC, 2009; HA-THUC, V.; MEJOVA, Y. et al., 2009; KIM; TIAN et al., 2009; MOON; KIM et al., 2009; WANG; ZHANG et al., 2009; YANG; SHI et al., 2009; CHEN; CHEN et al., 2010; HOLZ; TERESNIAK, 2010; STRÖTGEN; GERTZ, 2010; TANG; ZHANG, 2010). Muitas destas informações, apesar de serem consideradas um único e coerente bloco estático de dados, estão associadas, implícita ou explicitamente, a diferentes momentos no tempo. Por exemplo, bases de dados científicas normalmente possuem artigos que foram publicados ao longo de vários anos; mensagens de correio eletrônico são enviadas e recebidas diariamente; bases jornalísticas podem conter notícias publicadas diariamente, de várias épocas; sítios de notícias ou blogs podem ser atualizados a cada minuto, etc.

Vários autores têm discutido a importância de se considerar a dimensão *tempo* na análise de informações textuais. Segundo He et al. (2010), a explosão da Web trouxe uma enorme quantidade de informações, e assim criou uma demanda por novos meios de se gerenciar essa informação que está em constante *mudança*. Khy et al. (2008) afirmam que pesquisas relacionadas ao processamento de documentos que possuem uma ordem *temporal* são interessantes às áreas de recuperação e gestão da informação. Ha-Thuc et al. (2009) assinalam que *padrões temporais* descobertos podem revelar informações úteis sobre o comportamento dos diversos tópicos nos conjuntos de dados. He et al. (2009) afirmam que o entendimento de como tópicos na literatura científica *evoluem* é um interessante e importante problema. E de acordo com Alonso et al. (2009), na medida em que a quantidade de informação gerada aumenta rapidamente no mundo digital, o conceito de *tempo como uma dimensão* ao longo do

qual a informação pode ser organizada e explorada torna-se mais e mais importante.

Há também autores que citam importância da análise das relações entre elementos textuais no tempo. Segundo 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) afirmam que são necessários sistemas que mostrem como tópicos *emergem*, *modificam-se* e *desaparecem* (e talvez *reaparecem*) ao longo do tempo, e que técnicas de visualização são interessantes para mostrar os relacionamentos encontrados. Ha-Thuc et al. (2009) citam a importância de se explorar a *evolução* das *interações* entre comunidades em blogs. E Lin et al. (2009) afirmam que é necessário se considerar a dimensão tempo na análise de redes sociais.

Mais especificamente em relação à KDT e MT, Böttcher et al. (2008) afirmam que é necessário o emprego de uma perspectiva com orientação *temporal*, colocando o entendimento das *mudanças* no centro da descoberta de conhecimento. Já Baharudin et al. (2010) citam a *mineração de tendências* como uma oportunidade de pesquisa para a área de descoberta de conhecimento em dados não estruturados, por meio da aplicação algoritmos de MT para análise de tendências. Esses autores ainda afirmam que *fluxos* de textos requerem novos métodos e técnicas para gestão da informação. E segundo Wang et al. (2009), uma forma eficaz de se explorar a semântica bem como informação *temporal* em fluxos de textos é por meio de *mineração de tópicos*, o qual pode mais adiante facilitar outros procedimentos de *descoberta de conhecimento*.

Os trabalhos relativos à análise temporal de coleções de documentos textuais são basicamente divididos em Análise de Tendências (FELDMAN; DAGAN, 1995; LENT; AGRAWAL et al., 1997; FELDMAN; AUMANN et al., 1998; FELDMAN; DAGAN et al., 1998; MONTES-Y-GÓMEZ; GELBUKH et al., 2001), Detecção de Tendências Emergentes (KONTOSTATHIS; GALITSKY et al., 2004; MÖRCHEN; DEJORI et al., 2008; MÖRCHEN; FRADKIN et al., 2008; GOORHA; UNGAR, 2010), estudo de *burstness* (KLEINBERG, 2002; FUNG; YU et al., 2005; HE; CHANG et al., 2007; SUBASIC; BERENDT, 2008; 2010), Detecção e Rastreamento de Tópicos (ALLAN; PAPKA et al., 1998; ALLAN, 2002; MAKKONEN; AHONEN-MYKA et al., 2004; LI; WANG et al., 2005; ZHANG; ZI et

al., 2007; CHEN; CHEN et al., 2010; HOLZ; TERESNIAK, 2010; ROSSI; NEVILLE, 2010; YONGHUI; YUXIN et al., 2010), *Evolutionary Theme Patterns* (MEI; ZHAI, 2005; MEI; LIU et al., 2006; LIU; MERHAV et al., 2009; SUBAŠIĆ; BERENDT, 2010), Detecção de Desvios (FELDMAN; DAGAN, 1995; ARNING; RAGHAVAN, 1996; FELDMAN; AUMANN et al., 1998; KNORR; NG et al., 2000; MONTES-Y-GÓMEZ; GELBUKH et al., 2001; KAMARUDDIN; HAMDAN et al., 2007), Regras de Associação Temporais (LEE; LIN et al., 2001; NØRVÅG; ERIKSEN et al., 2006; BOUANDAS; OSMANI, 2007; GHARIB; NASSAR et al., 2010) e abordagens visuais (FELDMAN; AUMANN et al., 1998; HAVRE; HETZLER et al., 2002; SAGA; TSUJI et al., 2010; ŠILIĆ; DALBELO BAŠIĆ, 2010).

Apesar da existência desses trabalhos relativos à análise temporal de coleções de documentos textuais, geralmente são estudos específicos e não apresentam um Modelo para Descoberta de Conhecimento em Textos que seja independente de domínio e que permita o uso de diferentes algoritmos e técnicas de Mineração de Textos com ênfase nos relacionamentos e na dimensão tempo.

## 1.1 PROBLEMA DE PESQUISA

A partir do contexto acima mencionado, o seguinte problema é identificado:

Como descobrir padrões a partir de informações não estruturadas analisando a evolução dos relacionamentos entre os elementos textuais ao longo do tempo?

## 1.2 PRESSUPOSTOS DA PESQUISA

Considerando o problema acima mencionado os seguintes pressupostos da tese são apresentados:

- O aumento expressivo na quantidade de informação disponível demanda o desenvolvimento de modelos de engenharia do conhecimento para se desenvolver sistemas baseados em conhecimento que apoiem os diversos processos da gestão do conhecimento;

- Grande parte das informações disponíveis atualmente são não estruturadas e temporais. Estas duas características exigem que se desenvolvam modelos específicos para se lidar com essas informações;
- Há uma importância crescente em se desenvolver métodos para descoberta de conhecimento a partir da análise dos relacionamentos entre elementos textuais;
- Devido à rapidez com que se produz novas informações, o desenvolvimento de métodos para descoberta de conhecimento a partir da análise dos aspectos temporais destas informações torna-se relevante;
- Diversos materiais, métodos e ferramentas computacionais para processamento textual podem ser integradas e combinadas em um modelo de descoberta de conhecimento em fontes de informação não estruturadas com ênfase na evolução dos relacionamentos entre elementos textuais ao longo do tempo.

## 1.3 OBJETIVOS DO TRABALHO

### 1.3.1 Objetivo Geral

O objetivo geral desta tese é desenvolver um modelo de descoberta de conhecimento a partir de informações não estruturadas que possibilite analisar a evolução dos relacionamentos entre os elementos textuais ao longo do tempo.

### 1.3.2 Objetivos Específicos

Com a finalidade de atingir o objetivo geral, têm-se os seguintes objetivos específicos:

- Investigar e propor uma forma de se identificar e representar o peso dos relacionamentos diretos (correlação) e indiretos (associação) entre os elementos textuais ao longo do tempo;
- Identificar na literatura métodos, técnicas e algoritmos relativos à correlação, associação e análise temporal de informações

textuais, que possam ser utilizados na etapa de mineração textos do modelo proposto;
- Demonstrar a viabilidade do modelo proposto por meio do desenvolvimento de um protótipo e sua aplicação em um estudo de caso;
- Analisar as contribuições do modelo proposto à área de descoberta de conhecimento em textos por meio de uma análise comparativa com outros modelos existentes na literatura.

## 1.4 PRINCIPAIS CONTRIBUIÇÕES

Em resumo, apresenta-se abaixo as principais contribuições desta tese:

- O modelo de *Temporal Knowledge Discovery in Texts* (TKDT), baseado no modelo de KDT, com ênfase no aspecto temporal dos relacionamentos entre os elementos textuais. Trata-se de um modelo que estende dois modelos de KDT (MOONEY; NAHM, 2005; GONÇALVES, 2006), acrescentando novas noções, sendo a mais importante a dimensão temporal nos relacionamentos entre os elementos textuais;
- A etapa de *Temporal Text Mining* (TTM), o qual permite que os diversos algoritmos para análise temporal de informações textuais, em conjunto com técnicas de visualização e RI, sejam utilizados para apoiar os usuários em tarefas intensivas em conhecimento;
- Uma ontologia que representa conceitualmente as dimensões de análise do modelo, e o mapeamento dessa ontologia em um modelo dimensional de dados.
- Um modelo genérico para representação e análise de relacionamentos diretos e indiretos entre elementos textuais independentemente de domínio;
- A implementação de um protótipo e sua aplicação em um estudo de caso.

## 1.5 CONTEXTUALIZAÇÃO DO TRABALHO NO PROGRAMA

De acordo com Schreiber et al. (2002) e Studer et al. (2000), a *nova engenharia do conhecimento* tem por objetivo o desenvolvimento de métodos, técnicas e ferramentas que permitam que o conhecimento seja gerenciado e manipulado de maneira mais eficiente. Segundo Rautenberg (2009), quando esses métodos e técnicas são baseados em IA, eles podem ser denominados Agentes Computacionais da Engenharia do Conhecimento. Esse conceito é baseado na definição de agentes dada por Schreiber et al. (2002): agentes são indivíduos ou *sistemas computacionais* que, dado um domínio particular de interesse, são capazes de executar uma *tarefa intensiva em conhecimento*. Assim, o modelo de KDT proposto neste trabalho pode ser considerado um Agente Computacional da Engenharia do Conhecimento, passível de ser utilizado para auxiliar pessoas na execução de tarefas intensivas em conhecimento no contexto da GC.

Outro aspecto desta pesquisa que a contextualiza na área de Engenharia do Conhecimento está no fato de o modelo proposto prever o uso de conhecimento de domínio, que pode estar representado através de ontologias, tesauros, dicionários, etc.

O entendimento da adequação desta proposta de tese, de acordo com as três noções apresentadas acima, pode ser reforçada a partir da leitura do objeto de pesquisa e objetivo principal do Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento (EGC)[3]:

> O objeto de pesquisa do EGC refere-se aos macroprocessos de explicitação, gestão e disseminação do conhecimento. Estes incluem os processos de criação (e.g., inovação de ruptura), descoberta (e.g., redes sociais), aquisição (e.g., inovação evolutiva), formalização/codificação (e.g., ontologias), armazenamento (e.g., memória organizacional), uso (e.g., melhores práticas), compartilhamento (e.g., comunidades de prática), transferência (e.g., educação corporativa) e evolução (e.g., observatório do conhecimento) [...] Deste modo, o objetivo do EGC consiste em

[3] http://www.egc.ufsc.br/htms/vermais_index.htm

> investigar, conceber, desenvolver e aplicar modelos, métodos e técnicas relacionados tanto a processos/bens/serviços como ao seu conteúdo técnico-científico [...]

Como descrito acima no objetivo principal do EGC, um dos processos a serem pesquisados é a "descoberta de conhecimento". Logo, essa pesquisa está em consonância com os objetivos do EGC, pois se propõe um modelo de "descoberta de conhecimento" em textos, onde o elemento principal do modelo que se relaciona ao conhecimento como fator de produção é o atributo temporal das informações não estruturadas.

## 1.6 DELIMITAÇÃO DO TRABALHO

Como foi apresentado anteriormente, o modelo proposto prevê o uso de conhecimento de domínio nas análises que pode estar contido em ontologias, dicionários, tesauros, etc. A forma de se obter esse conhecimento fica fora do escopo desta pesquisa. É uma tarefa que depende do caso concreto e fica sob responsabilidade dos usuários do modelo. Outro aspecto que se deve levar em consideração é fato de que a qualidade do conhecimento de domínio interfere nos resultados das análises.

Em relação à dimensão tempo do modelo, está fora do escopo desta pesquisa o estudo de meios de obtenção da data dos documentos. É responsabilidade de aplicações concretas e domínios específicos obtenção dessa informação.

E apesar de métodos, técnicas e ferramentas de visualização de informação serem importantes para apresentação dos resultados obtidos por meio da aplicação do modelo, está fora do escopo desse trabalho o estudo dessa área. Quando necessário, serão utilizadas ferramentas de visualização já disponíveis para esse propósito.

## 1.7 MÉTODO DE PESQUISA

Para atingir os objetivos desta pesquisa, o trabalho foi dividido nas seguintes etapas:

- Revisão da literatura científica relevante para o desenvolvimento deste trabalho: (a) conceitos de Engenharia do Conhecimento (b) KDT/MT, como o arcabouço que suporta o modelo proposto nesse trabalho; (c) correlação de elementos textuais, que permite a identificação de relacionamentos diretos entre elementos textuais; (d) associação de elementos textuais, a partir dos trabalhos em DBL, permite a identificação dos relacionamentos indiretos entre elementos textuais; (e) análise temporal de informações textuais; (f) modelos baseados em coocorrências, para cálculo de correlações e construção dos vetores de contexto; e (g) RI, para identificar formas eficientes de representação e manipulação de informações textuais, Modelo Espaço Vetorial (MEV) para representação dos vetores de contexto dos elementos textuais, e medidas de similaridades entre vetores para cálculo da associação;
- Especificação do modelo de TKDT detalhando-se todos os seus componentes e como esses atingem os objetivos do trabalho quando integrados;
- Demonstração de viabilidade por meio do desenvolvimento de um protótipo baseado no modelo proposto e sua aplicação em um estudo de caso;
- Análise das contribuições do modelo proposto à área de descoberta de conhecimento em textos por meio de uma análise comparativa com outros modelos de descoberta de conhecimento em textos;
- Discussão das conclusões obtidas e dos possíveis trabalhos futuros para aprimoramento do modelo proposto.

## 1.8 ORGANIZAÇÃO DO TRABALHO

Este trabalho é composto de cinco capítulos, sendo os demais descritos a seguir.

- Capítulo 2. Fundamentação Teórica: neste capítulo apresentam-se as áreas de Engenharia do Conhecimento, Descoberta de Conhecimento em Textos, Mineração de Textos, correlação e associação de elementos textuais, análise temporal de informações textuais, modelos baseando em coocorrência,

representação vetorial, similaridade de vetores e estrutura de índice-invertido;

- Capítulo 3: Modelo Proposto: neste capítulo apresenta-se o modelo de TKDT proposto, discutindo-se cada parte do modelo em detalhes e como esses atingem os objetivos do trabalho quando integrados;
- Capítulo 4: Protótipo Baseado no Modelo Proposto: capítulo que mostra a viabilidade do modelo proposto por meio da implementação de um protótipo de sistema baseado em conhecimento a partir do modelo proposto. É apresentada a arquitetura do protótipo, seus módulos e suas relações com o modelo;
- Capítulo 5: Demonstração de Viabilidade e Análise Comparativa: capítulo que apresenta um estudo de caso no qual o protótipo desenvolvido é aplicado em um conjunto de informações. Também são discutidas as contribuições do modelo à área de descoberta de conhecimento em textos por meio de uma análise comparativa com outros dois modelos de KDT;
- Capítulo 6: Conclusões e Trabalhos Futuros: este capítulo descreve as conclusões obtidas com essa pesquisa e apresenta algumas sugestões de possíveis trabalhos futuros;

# 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta em suas seções o referencial teórico do modelo proposto. Está dividido em quatro partes: engenharia do conhecimento, descoberta de conhecimento em textos, modelos baseados em coocorrências e recuperação de informação. Na primeira, discutem-se alguns conceitos da área de Engenharia do Conhecimento (EC), e são apresentadas algumas definições relativas ao contexto desta tese. Na segunda parte, é apresentada a área de Descoberta de Conhecimento em Banco de Dados – KDD (*Knowledge Discovery in Databases*) e Mineração de Dados (MD), e suas especializações para fontes de informação textuais: Descoberta de Conhecimento em Textos – KDT (*Knowledge Discovery in Text*) e Mineração de Textos (MT). Além disso, se discutirá os aspectos de KDT e MT que tem relação direta com os objetivos deste trabalho: correlação de elementos textuais (análise de relacionamentos diretos); associação de elementos textuais (análise de relacionamentos indiretos) entre elementos textuais; e análise temporal de informações textuais. A terceira parte apresenta em detalhes alguns dos principais modelos baseados em coocorrência. Esses modelos são utilizados para o cálculo da correlação e associação de elementos textuais. Na quarta e última parte são apresentados conceitos da área de recuperação de informação que utilizados neste trabalho: representação vetorial e similaridade de vetores, que são utilizados para o cálculo da associação entre elementos textuais; e a estrutura de índice invertido, utilizada para manipular eficientemente informações textuais.

## 2.1 ENGENHARIA DO CONHECIMENTO

A Engenharia do Conhecimento (EC) se desenvolveu a partir do final da década de 70 voltada à construção de Sistemas Baseados em Conhecimento (SBC) dentro da área de Inteligência Artificial (IA) (SCHREIBER; AKKERMANS et al., 2002). Mais recentemente, a construção de SBCs se tornou uma atividade complexa, devido ao surgimento da Gestão do Conhecimento (GC) e dos avanços das Tecnologias da Informação e Comunicação (TICs). Deste modo, a EC evoluiu para a área responsável por métodos e ferramentas para a construção sistêmica e controlada de SBCs para apoiar os diversos

processos (criação, organização, formalização, compartilhamento, aplicação e refinamento) da GC (HENDRIKS, 1999; STUDER; DECKER et al., 2000; SCHREIBER; AKKERMANS et al., 2002; NISSEN, 2006). Essa visão é atualmente difundida na comunidade científica, que pontua que a EC se refere a todos os aspectos técnicos, científicos e sociais envolvidos na construção, manutenção e uso de SBCs (KOED, 2009).

Segundo Deng e Yu (2006), apesar de muitas metodologias e técnicas de EC se preocuparem com os ativos de conhecimento de uma organização, ainda existem desafios a considerar, tais como a preparação e a estruturação do conhecimento. Nesse sentido, Schreiber et al. (2002) sugerem o CommonKADS (Figura 1) como uma metodologia de EC para modelagem de SBCs para a GC.

O CommonKADS une as dimensões pessoas, processos, conteúdo e tecnologia a seus modelos de Organização, Tarefas, Agentes, Conhecimento, Comunicação e Projeto. Assim, os SBCs modelados de acordo com o CommonKADS consideram a GC em nível de contexto, de conceito e de artefato de seus modelos. O uso do CommonKADS está em consonância com o que é descrito por Cheung (2006), o qual afirma que um SBC é modelado segundo técnicas reutilizáveis de representação e extração de conhecimento.

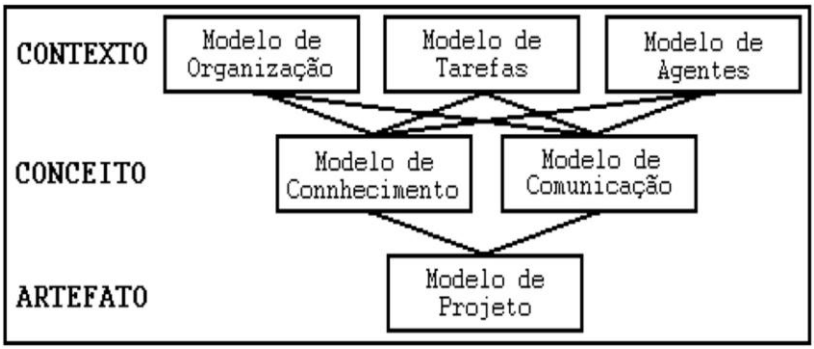


Figura 1 – Modelos do CommonKADS.
Fonte: adaptado de (SCHREIBER; AKKERMANS et al., 2002)

### 2.1.1 Dado, Informação e Conhecimento

Schreiber et al. (2002) apresentam as seguintes definições para dado, informação e conhecimento que, segundo esses mesmos autores, fornecem uma demarcação sobre a qual há consenso na literatura:

Dados são milhares de *sinais* não interpretados que alcançam nossos sentidos a cada minuto. Por exemplo, uma luz vermelha, verde ou amarela de um semáforo num cruzamento. Computadores são repletos de dados: sinais que consistem em números, caracteres e outros símbolos que são mecanicamente manipulados em grandes quantidades.

Informação é dado provido com *significado*. Para um motorista de carro, uma luz vermelha de um semáforo não é apenas um sinal de algum objeto colorido, e sim uma indicação para parar. Em contraste, um ser alienígena que acabou de chegar à Terra provavelmente não atribuirá o mesmo significado à luz vermelha. O dado é o mesmo, mas a informação é diferente.

Conhecimento é o conjunto de dados e informações que as pessoas levam para auxiliar em algum uso prático em *ação*, para executar tarefas e criar informação nova. O conhecimento acrescenta dois aspectos diferentes: (a) um senso de *propósito*, visto que o conhecimento é a *intellectual machinery* usada para alcançar uma meta; e (b) uma *capacidade generativa*, porque umas das maiores funções do conhecimento é produzir novas informações. É por isso que o conhecimento é dito ser um novo "fator de produção".

### 2.1.2 Tarefas

De acordo com Schreiber et al. (2002), no contexto da EC, tarefa é algum trabalho que precisar ser feito por um agente. Nesta tese, o maior interesse está nas tarefas "intensivas em conhecimento", que são tarefas nas quais o conhecimento desempenha algum papel importante.

### 2.1.3 Agentes

No contexto da EC, agentes são indivíduos ou sistemas computacionais que são capazes de executar uma tarefa em certo domínio particular de interesse (SCHREIBER; AKKERMANS et al.,

2002). Quando se restringe aos sistemas computacionais, estes também podem ser chamados de Agentes Computacionais da Engenharia do Conhecimento (RAUTENBERG, 2009). E de acordo com Huang (2009), esses agentes são projetados em função de alguma tarefa de resolução de problemas via combinação de métodos e técnicas de IA e bases de conhecimento específicas. Esse aspecto enfatiza a importância de agentes computacionais da EC diante a GC na execução e/ou auxílio em tarefas intensivas em conhecimento. No contexto desta tese, a área de Descoberta de Conhecimento em Textos, abordada na seção 2.2, é um exemplo de Agente Computacional da Engenharia do Conhecimento.

### 2.1.4 Informações não Estruturadas

Apesar de um texto em linguagem natural ser estruturado no sentido de possuir uma estrutura sintática, a referência a "estrutura" é feita no âmbito da Ciência da Computação. Os dados ditos "estruturados" estão em bancos de dados – identificados, indexados e armazenados em registros e campos específicos. Dados "semiestruturados" possuem marcação com tags em linguagem XML. Textos em e-mails, relatórios, artigos etc., nesse sentido, são considerados dados "não estruturados". No contexto deste trabalho, não se diferencia "semiestruturados" de "não estruturados", sendo este último termo usado preferencialmente.

As informações não estruturadas utilizadas nesta tese são normalmente organizadas em documentos. Um documento pode ser definido como uma unidade discreta de informação textual que usualmente está relacionada com algum documento do mundo real, tal como um relatório, um memorando, um e-mail, um artigo, etc. Uma coleção de documentos é aqui chamada de fonte de informação.

As fontes de informação utilizadas no modelo proposto são caracterizadas por possuírem algum atributo temporal como, por exemplo, a data de criação dos documentos. Assim, essas fontes se caracterizam pela inclusão de novos documentos e/ou atualização dos documentos existentes ao longo do tempo.

## 2.2 DESCOBERTA DE CONHECIMENTO EM TEXTOS

Dentro deste contexto de aumento expressivo na quantidade e na importância da informação para as organizações, a área de descoberta de conhecimento desempenha um papel cada vez mais importante. Como afirmam Hair et al. (1998), a área de descoberta de conhecimento se baseia nessa avalanche de informações como também em questionamentos sobre essa informação. Assim, a análise de dados passa a ter uma caráter mais exploratório, visando identificar ou explicitar conhecimento oculto em fontes de informação. Essa tarefa é tradicionalmente de responsabilidade da área de Descoberta de Conhecimento em Bases de Dados (KDD). Segundo Fayyad (1996), KDD é um processo não trivial de identificação, a partir de dados, de padrões novos, válidos, potencialmente úteis e compreensíveis. Nessa definição, os dados representam um conjunto de fatos, e um padrão é uma expressão em alguma linguagem que descreve um subconjunto de dados ou um modelo aplicável a esse subconjunto. Portanto, em KDD extrair um padrão consiste na atividade de adaptar um modelo aos dados ou descobrir alguma estrutura neles; ou, de maneira geral, encontrar alguma descrição de alto nível em um conjunto de dados.

O termo "processo" implica que KDD é composto de vários passos (Figura 2), os quais envolvem preparação dos dados, busca por padrões, avaliação do conhecimento e refinamento, que são repetidos em múltiplas iterações. Por "não trivial" entende-se que envolve alguma busca ou inferência e que não é apenas uma computação direta de valores predefinidos. Os padrões descobertos devem ser válidos perante os novos dados, com algum grau de certeza. Também é desejável que esses padrões sejam novos e potencialmente úteis. Isso quer dizer que eles devem trazer algum benefício para o usuário. Por último, os padrões devem ser compreensíveis. Se isso não for possível imediatamente, devem ser alvo, então, de algum método de pós-processamento. Na Figura 2 tem-se uma visão geral do processo de KDD, o qual envolve a seleção, o pré-processamento, a transformação do dado, a utilização de algoritmos especializados e a geração de conhecimento (FAYYAD, 1996). O modelo possui processos repetitivos entre as fases, isto é, a cada avaliação da fase atual, a(s) fase(s) anterior(es) pode(m) sofrer ajuste(s).

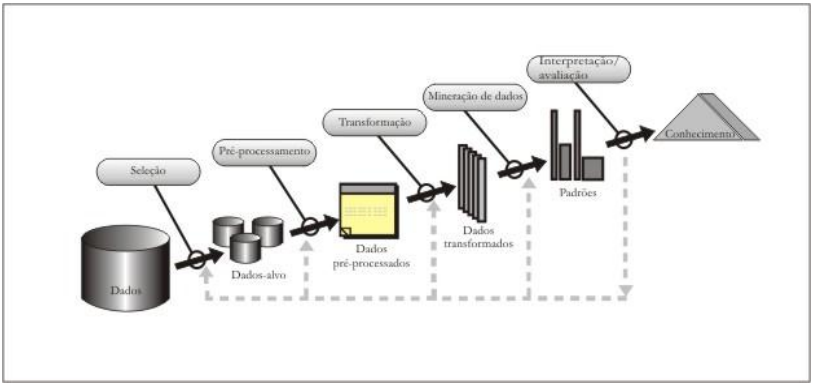


Figura 2 – Uma visão geral do processo de KDD.
Fonte: adaptado de (FAYYAD, 1996)

Como pode ser visto na figura, trata-se de um processo repetitivo no qual todos os passos são importantes para se atingir o objetivo de descoberta de conhecimento. Deve ser visto como um método iterativo, e não como uma ferramenta de análise automática (MANNILA, 1996). Um dos seus principais passos é a Mineração de Dados (MD), sendo responsável pela aplicação de algoritmos com o propósito de identificar padrões em uma base de dados (FAYYAD; PIATETSKY-SHAPIRO et al., 1996a). A MD pode ser entendida como "a exploração e a análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados, com o objetivo de descobrir padrões e regras significantes" (BERRY; LINOFF, 1997).

Quanto às metas da MD, Fayyad et al. (1996b) apresentam dois tipos: verificação, em que o sistema é limitado a confirmar as hipóteses do usuário (teste de hipóteses); e descoberta, em que o sistema automaticamente encontra novos padrões. A descoberta é ainda dividida em: (1) previsão, etapa em que o sistema procura padrões para a proposta de predição de comportamento futuro de algumas entidades (parte de diversas variáveis para prever outras variáveis ou valores desconhecidos); e (2) descrição, etapa em que o sistema procura por padrões com a proposta de apresentá-los ao usuário de forma compreensível.

Com a crescente aumento da quantidade de informações textuais (DÖRRE; GERSTL et al., 1999; TAN, 1999; LYMAN, 2003; HIMMA, 2007) houve a necessidade de se adaptar os métodos tradicionais de

descoberta de conhecimento para se lidar com dados não estruturados, surgindo assim, a Descoberta de Conhecimento em Textos (KDT) e a Mineração de Textos (MT).

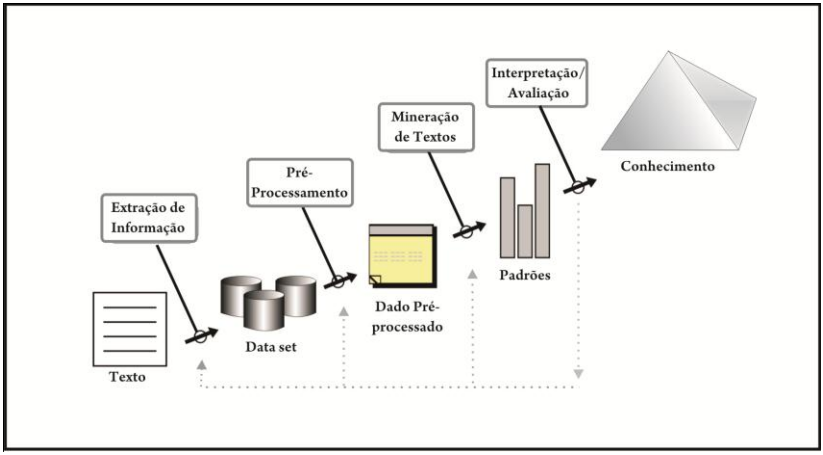


Figura 3 – Modelo de Descoberta de Conhecimento em Textos (KDT).
Fonte: Adaptado de (MOONEY; NAHM, 2005)

Análogo ao KDD, o KDT, que é apresentado na Figura 3, refere-se ao processo de maneira geral, enquanto que MT pode ser vista como uma extensão da Mineração de Dados tradicional. A MT representa o processo de extração de padrões relevantes e não triviais a partir de bases de dados semi ou não estruturadas (FELDMAN; DAGAN, 1995; FELDMAN; FRESKO et al., 1998; DÖRRE; GERSTL et al., 1999; WITTEN; BRAY et al., 1999; NASUKAWA; NAGANO, 2001; MOONEY; NAHM, 2005; GUPTA; LEHAL, 2009). Utiliza principalmente de conceitos de PLN, RI, EI e da estatística. Tarefas típicas de MT incluem classificação e agrupamento de textos, extração de entidades/conceitos, de sumarização de documento, análise de relacionamentos, descoberta de regras, etc.

Gonçalves (2006) apresenta um modelo de KDT (Figura **4**) baseado na correlação de elementos textuais e expansão vetorial. O objetivo desse modelo é descobrir relacionamentos latentes entre elementos textuais e, assim, promover melhoramentos na representação de documentos e fornecer suporte a aplicações de engenharia e gestão do conhecimento.

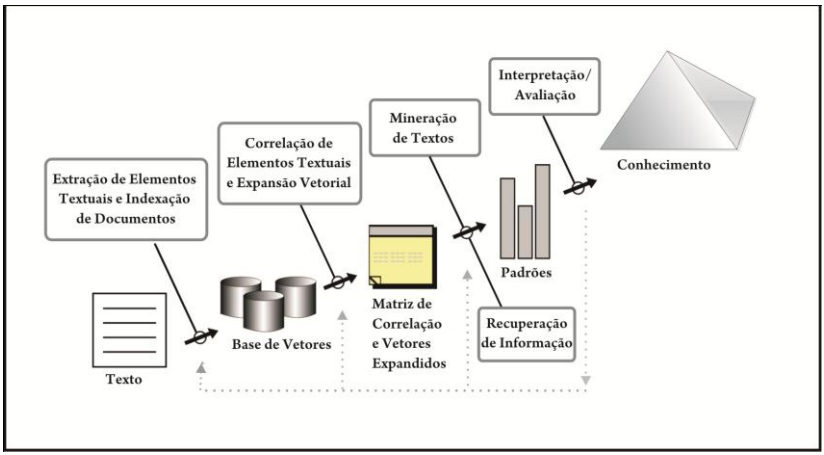


Figura 4 – Modelo de KDT baseado na correlação de elementos textuais e expansão vetorial.
Fonte: adaptado de (GONÇALVES, 2006)

Nas seções seguintes são apresentados alguns conceitos e pesquisas em KDT estão relacionados ao principais componentes do modelo proposto neste trabalho: a) correlação de elementos textuais (análise de relacionamentos diretos); b) associação de elementos textuais (análise de relacionamentos indiretos); e c) análise temporal de informações textuais.

### 2.2.1 Correlação de Elementos Textuais

Várias pesquisas em MT estão voltadas à análise de relacionamentos diretos entre conceitos em informações textuais (ZHU; GONÇALVES et al., 2005; ERHARDT; SCHNEIDER et al., 2006; GONÇALVES; ZHU et al., 2006; GONÇALVES; BEPPLER et al., 2006; ZHU; GONÇALVES et al., 2007; DAVIDOV; RAPPOPORT, 2008; GARTEN; ALTMAN, 2009; YAN; MATSUO et al., 2009; BUI; NUALLAIN et al., 2010; CHEN, 2010; MESQUITA; MERHAV et al., 2010; ROSE; ENGEL et al., 2010; SÁNCHEZ, 2010). Para se analisar tais relacionamentos, a forma mais simples consiste em contar o número

de coocorrências entre os elementos textuais. Contudo, existem métodos mais sofisticados que permitem determinar o *peso* do relacionamento.

Dentre esses modelos[4], têm-se alguns com origem na estatística descritiva, tais como o teste *t*, *Chi-square* $x^2$ e o *Z score* (MANNING; SCHÜTZE, 1999); os de origem na teoria da informação, tais como Informação Mútua (IM) (CHURCH; HANKS, 1990; CHURCH; GALE, 1991) e *Phisquared* ($\Phi^2$) (CHURCH; GALE, 1991); os modelos com base mais empírica como, por exemplo, o algoritmo CORDER (ZHU; GONÇALVES et al., 2005) e o *Latent Relation Discovery* (LRD) (GONÇALVES; ZHU et al., 2006); e tem-se ainda o modelo Indexação de Semântica Latente (ISL), que tem por objetivo capturar a estrutura semântica de coleções de documentos através da correlação de termos e documentos (DEERWESTER; DUMAIS et al., 1990; DING, 2000). Alguns desses modelos já foram utilizados no contexto da engenharia e gestão do conhecimento (GONÇALVES; BEPPLER et al., 2006). A análise dos relacionamentos *diretos* entre elementos textuais com base em suas coocorrências é chamada de *correlação* de elementos textuais.

A maior parte dos trabalhos relacionados à correlação de elementos textuais tem sido feita no contexto de *Biomedical Text Mining* (COHEN; HERSH, 2005; ERHARDT; SCHNEIDER et al., 2006; DAI; CHANG et al., 2010). Nessa área, busca-se analisar os relacionamentos entre entidades biológicas tais como genes, proteínas, doenças, drogas, químicos, etc. Várias ferramentas de MT analisam as coocorrências entre essas entidades dentro de sentenças, parágrafos, etc., para construir, por exemplo, redes do tipo "proteína-proteína", "gene-droga" e "droga-doença" (PEREZ-IRATXETA; BORK et al., 2001; CHANG; ALTMAN, 2004; CHEN; SHARP, 2004; ALAKO; VELDHOVEN et al., 2005; HOFFMANN; VALENCIA, 2005; PLAKE; SCHIEMANN et al., 2006; REBHOLZ-SCHUHMANN; KIRSCH et al., 2007; TSURUOKA; TSUJII et al., 2008; GARTEN; ALTMAN, 2009; THEOBALD; SHAH et al., 2009; BARBOSA-SILVA; SOLDATOS et al., 2010; BUI; NUALLAIN et al., 2010; GARTEN; TATONETTI et al., 2010).

O estudo de correlação de elementos textuais também tem sido utilizado fora do contexto das ciências biomédicas. Por exemplo, Mesquita et al. (2010) apresentam um sistema para extração de redes

4 Na seção 2.3 deste capítulo será apresentada uma descrição detalhada de alguns dos principais modelos baseados em coocorrência.

sociais a partir da *blogosfera*. Esse sistema, chamado de *SOcial Network Extraction* (SONEX), identifica entidades (pessoas, organizações, entidades geopolíticas, etc.) e extrai relacionamentos entre elas a partir das coocorrências de entidades em uma mesma sentença. Outro exemplo é uso de correlação de elementos textuais na análise de registros de buscas (*log search engines*) com o objetivo de sugerir termos relacionados ao termo que o usuário digitou no campo de busca, com base em buscas feitas anteriormente por outros usuários (CHEN, 2010). Há também trabalhos que envolvem a coocorrência de termos na Web. Neste sentido, Sánchez (2010) apresenta uma metodologia que utiliza análise estatística das coocorrências entres termos na Web para descobrir atributos de conceitos de uma ontologia. E também o trabalho de Turney (2004), que utiliza a mesma ideia para a tarefa de *Word Sense Disambiguation*.

Há ainda trabalhos relacionados à indução automática de taxonomias (YANG; CALLAN, 2009); descoberta do *tipo* de relação entre duas entidades (DAVIDOV; RAPPOPORT, 2008; YAN; MATSUO et al., 2009); *Sentiment Analysis* (TURNEY; LITTMAN, 2002; YU; HATZIVASSILOGLOU, 2003; TANG; TAN et al., 2009); extração automática de palavras-chave de documentos individuais (ROSE; ENGEL et al., 2010); e à *Abbreviation Recognition* (LIU; FRIEDMAN, 2003; OKAZAKI; ANANIADOU, 2006; ZHOU; TORVIK et al., 2006), que consiste na identificação de formas expandidas de abreviações como, por exemplo, identificar o termo "Inteligência Artificial" para a abreviação IA.

### 2.2.2 Associação de Elementos Textuais

Uma área de pesquisa que tem por objetivo encontrar relacionamentos indiretos em fontes de informação textuais é a Descoberta Baseada em Literatura (DBL). O seu objetivo é a aplicação de métodos de MT para a descoberta de novos conhecimentos a partir dos relacionamentos indiretos entre elementos textuais presentes na literatura científica. A análise dos relacionamentos *indiretos* entre elementos textuais, com base nos contextos nos quais eles aparecem nos documentos, é chamada de *associação* de elementos textuais.

Ela surgiu com o trabalho que Swanson fez com bases de artigos da área de Ciências Biomédicas (SWANSON, 1986). Em sua primeira

investigação, Swanson buscava informações sobre a Doença de Raynaud (*Raynaud's Disease*) – uma condição que resulta em restrição intermitente do fluxo sanguíneo para os dedos, disparado pelo frio ou estímulos emocionais (SWANSON, 1986; 1990; GORDON; LINDSAY, 1996). Na época dessa pesquisa, a cura para esta condição ainda não tinha sido encontrada. Apesar de Swanson não saber exatamente o que estava procurando, a sua revisão da literatura sobre o assunto resultou na descoberta de uma intervenção médica para a Doença de Raynaud. Ele descobriu isso através de relacionamentos indiretos contidos na literatura analisada. Primeiro, analisando a literatura sobre a "Doença de Raynaud", Swanson conseguiu fazer a conexão entre essa doença e o termo "Alta Viscosidade do Sangue" (*High Blood Viscosity*). Na revisão da literatura sobre "Alta Viscosidade do Sangue", ele encontrou uma conexão entre esse termo e o termo "Óleo de Peixe" (*Oil Fish*). Isto conduziu para a nova hipótese que "Óleo de Peixe" pode ser uma dieta suplementar útil para ajudar a diminuir a "Alta Viscosidade do Sangue" em seres humanos e então aliviar os sintomas da "Doença de Raynaud" (SWANSON, 1986). Assim, ele conseguiu achar um relacionamento indireto entre o termo "Doença de Raynaud" e o termo "Óleo de Peixe", através do termo "Alta Viscosidade do Sangue" (veja a Figura 5). Tal hipótese foi posteriormente testada e comprovada por pesquisadores da área médica.

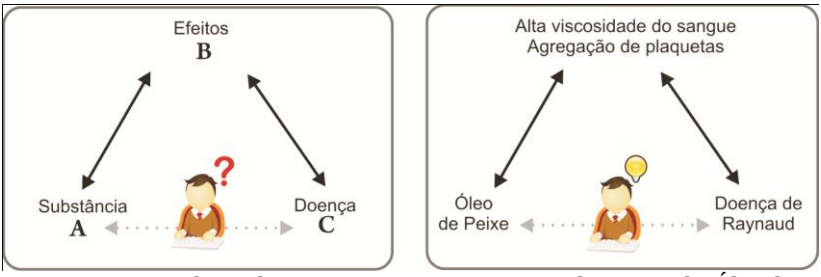


Figura 5 – Descoberta de Swanson: conexão "Doença de Raynaud - Óleo de Peixe".
Fonte: adaptado de (WEEBER; KLEIN et al., 2001)

A área de DBL surgiu devido ao enorme crescimento da quantidade de conhecimento científico durante o século passado (WEEBER, 2003). Uma das características do aumento de conhecimento científico é que cada cientista tem que interpretar grandes

quantidades de conhecimentos existentes e adquirir certas habilidades antes que eles possam contribuir para o seu domínio de conhecimento com a descoberta de conhecimento novo (WEEBER, 2003). Além disso, acompanhar os mais recentes desenvolvimentos para integrar novos conhecimentos a sua própria pesquisa não é uma tarefa simples para um cientista. Simon et al. (1997) afirmam que publicações científicas são o principal instrumento para acumulação e coordenação do conhecimento científico. Swanson (1986) mostrou que é possível usar essas publicações científicas para gerar conhecimento novo no contexto de DBL.

A premissa dessa abordagem é que há duas partes ou estruturas de conhecimento científico que não se comunicam entre si. Contudo, partes do conhecimento de uma dessas estruturas podem complementar o conhecimento da outra. Suponha que a comunidade científica sabe que B é uma das características da doença C. Outro grupo científico (disciplina ou estrutura de conhecimento) tem encontrado que a substância A afeta B. Descoberta, neste caso, é fazer o a ligação implícita AC através da conexão B (como no caso apresentado anteriormente). A Figura 6 ilustra esta situação.

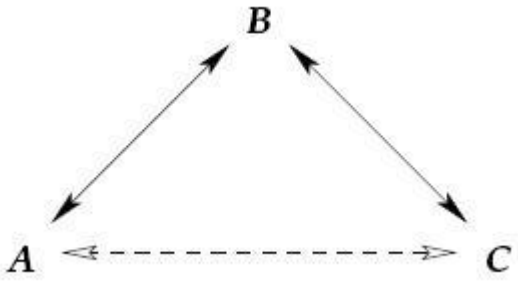


Figura 6 – Modelo ABC de Descoberta. Os relacionamentos AB e AC são conhecidos e relatados na literatura. O relacionamento implícito AC é uma suposta nova descoberta.
Fonte: (WEEBER, 2003)

Weeber et al. (2001) definiram duas abordagens de descoberta no modelo ABC: fechada e aberta. A descoberta *fechada* começa com A e C conhecidos. Podem ser uma associação observada, ou uma já hipótese já gerada. A descoberta nesta situação concentra-se em encontrar novos

Bs que podem explicar essa observação. O processo descoberta *aberta* inicia com a estrutura de conhecimento na qual o cientista participa (A). O primeiro passo é encontrar potenciais conexões B. Estes provavelmente serão encontrados dentro do próprio domínio. O passo crucial, contudo, é de B para C o qual é mais provável de estar fora do escopo do cientista, e pode então estar em qualquer ponto do espaço de conhecimento da ciência.

Desde 1988, Swanson tem usado ferramentas computacionais de análise textual para ajudá-lo no processo de estudo da literatura. Estas ferramentas evoluíram para uma ferramenta de suporte à descoberta chamada Arrowsmith (SWANSON; SMALHEISER, 1997; SMALHEISER; TORVIK et al., 2009). No contexto da ciências biomédicas, tem-se ainda os trabalhos de Gordon e Lindsay (1996), que usaram uma abordagem com princípios mais analíticos baseada em estatística de frequência de palavras; Lindsay e Gordon (1999), que usaram Trigramas e Análise Contextual; Gordon e Dumais (1998), com ISL; Weeber et al. (2001), os quais desenvolveram a ferramenta Literaby, que trabalha com conceitos ao invés de palavras/termos; Pratt e Yetisgen-Yildiz (2003), com a ferramenta LitLinker; Srinivasan (2004), com geração de hipóteses a partir do MEDLINE; Van der Eijk et al. (2004) com os *Associative Concept Spaces*; Wren et al. (2004), que utilizam modelos aleatórios; Hristovski et al. (2005), Kastrin e Hristovski (2008), com o software BITOLA; e van Haagen et al. (2009) que utilizam o *Concept Profile Method* para descoberta de interações entre proteínas a partir de Medline, implementado na ferramenta Nermal; e Gandra et al. (2010), que propõem uma metodologia para identificar e validar associações implícitas entre proteínas, que são descobertas através do sistema BioMAP (GANDRA; PRADHAN et al., 2003). Há ainda outros trabalhos que relacionados à descoberta de associações indiretas entre genes, químicos, doenças, etc. (COHEN, 2008; COHEN; SCHVANEVELDT et al., 2009; JORGE-BOTANA; OLMOS et al., 2009; PETRIC; URBANCIC et al., 2009; BAKER; HEMMINGER, 2010; COHEN; SCHVANEVELDT et al., 2010; ZHOU; PENG et al., 2010).

Há também alguns trabalhos de DBL fora do contexto das ciências biomédicas, como o trabalho de Cory (1997), que realizou um estudo com o objetivo de desenvolver uma metodologia para acelerar a pesquisa na área de Humanidades através da descoberta de analogias significantes que estejam latentes bases de artigos da área; e Gordon et

al. (2002), que realizaram alguns experimentos na área de Ciência da Computação, utilizando informações publicadas na *World Wide Web*, para encontrar novas áreas de aplicação para tecnologias existentes.

### 2.2.3 Análise Temporal de Informações Textuais

Em muitos domínios de aplicação encontram-se documentos textuais com alguma marcação de tempo (*timestamp*) associada. Por exemplo, notícias sobre um determinado assunto (dia da publicação), artigos científicos de uma área (ano da publicação), mensagens de e-mails (dia do envio ou recebimento), etc. Em tais informações podem haver padrões temporais interessantes. Por exemplo, um evento coberto nos artigos de notícias geralmente têm uma estrutura temporal e evolucionária consistindo de temas (subtópicos) que caracterizam o começo, progresso, e impacto do evento. Por exemplo, no caso de artigos científicos, o estudo de um tópico em algum período de tempo pode ter influenciado ou estimulado o estudo de outro tópico em outra época posterior (MEI; ZHAI, 2005). Assim, esse tipo de análise permite ao usuário encontrar similaridades e diferenças nas bases textuais entre os períodos de tempo de uma maneira que facilite ver a variação da importância dos conceitos e seus relacionamentos ao longo do tempo.

Dentro desse contexto, a Análise de Tendências (AT) é o termo geralmente usado para descrever a análise da distribuição de conceitos através de múltiplos subconjuntos de documentos no tempo (FELDMAN; DAGAN, 1995; LENT; AGRAWAL et al., 1997; FELDMAN; AUMANN et al., 1998; FELDMAN; DAGAN et al., 1998; MONTES-Y-GÓMEZ; GELBUKH et al., 2001). O trabalho sobre AT de Montes-y-Gómez et al. (2001) busca analisar textos de notícias para responder a perguntas tais 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? Quais são os assuntos que estão emergindo ou desaparecendo? Entre outras.

Muitos dos trabalhos em AT são chamados de Detecção de Tendências Emergentes (DTE) (KONTOSTATHIS; GALITSKY et al., 2004; MÖRCHEN; DEJORI et al., 2008; MÖRCHEN; FRADKIN et al., 2008; GOORHA; UNGAR, 2010). Segundo tais autores, uma tendência emergente é um assunto (tópico, área, etc.) que está crescendo em

interesse e utilidade ao longo do tempo. Por exemplo, XML emergiu como uma tendência no meio dos anos 90. A Tabela 1 mostra os resultados de uma busca em uma base bibliográfica da área de Ciência da Computação utilizando-se a palavra "XML". Como pode ser visto, XML emergiu a partir de 1994 a 1997 e em 1998 estava bem representada como um tópico na área de Ciência da Computação. Existem também trabalhos na área de DTE aplicados ao domínio de patentes, com o objetivo de se desenvolver ferramentas analíticas para o reconhecimento de *tecnologias* emergentes (POTTENGER; YANG, 2001; AHMAD; AL-THUBAITY, 2003; YOON; PARK, 2004; KIM; SUH et al., 2008; KIM; TIAN et al., 2009).

Outros trabalhos relacionados a AT e a DTE envolvem o estudo de *burstness* em fluxos de documentos (*document streams*). Trata-se da descoberta de tópicos que possuem uma alta frequência em certo período de tempo, ganhando volume rapidamente no começo desse período e (usualmente) desaparecendo na mesma velocidade (KLEINBERG, 2002; FUNG; YU et al., 2005; HE; CHANG et al., 2007; SUBASIC; BERENDT, 2008; 2010).

| **Ano** | **Número de Documentos** |
|---|---|
| 1994 | 3 |
| 1995 | 1 |
| 1996 | 8 |
| 1997 | 10 |
| 1998 | 170 |
| 1999 | 371 |

Tabela 1 – Emergência de XML no meio dos anos 90, segundo resultado de busca em base bibliográfica da área de Ciência da Computação.
Fonte: (KONTOSTATHIS; GALITSKY et al., 2004)

Há também uma área de pesquisa chamada de Detecção e Rastreamento de Tópicos (DRT) (ALLAN; PAPKA et al., 1998; ALLAN, 2002; MAKKONEN; AHONEN-MYKA et al., 2004; LI; WANG et al., 2005; ZHANG; ZI et al., 2007; CHEN; CHEN et al.,

2010; HOLZ; TERESNIAK, 2010; ROSSI; NEVILLE, 2010; YONGHUI; YUXIN et al., 2010). Trata-se do desenvolvimento de métodos para detectar um tópico e rastreá-lo no tempo. O foco do DRT está em eventos descritos em textos de notícias: divide-se o texto em histórias coesas, localiza-se algum evento previamente não relatado, segue-se o desenvolvimento de tal evento, agrupando as notícias que discutem o mesmo evento. Um evento é alguma coisa que acontece em determinado tempo e lugar, o qual pode ser reportado consecutivamente por muitas notícias durante um período de tempo. Segundo He et al. (2010), como podem haver muitos documentos (notícias) que relatam a mesma informação (evento), torna-se importante que esse processo inclua a sumarização automática dos documentos, produzindo um conteúdo com as principais informações sobre o tópico. De maneira correlata, há trabalhos que envolvem *Evolutionary Theme Patterns*, os quais usam modelos probabilísticos para a descoberta, extração e a sumarização dos padrões de *evolução* de temas em bases textuais temporais (MEI; ZHAI, 2005; MEI; LIU et al., 2006; LIU; MERHAV et al., 2009; SUBAŠIĆ; BERENDT, 2010).

Outro tipo de análise temporal chama-se *Ephemeral Associations* (MONTES-Y-GÓMEZ; GELBUKH et al., 2001). Trata-se de uma tipo de análise que se permite ver a influência dos conceitos mais frequentes em um período sobre outros conceitos no mesmo período. Uma *Ephemeral Association* pode ser de dois tipos: inversa ou direta. Na associação inversa, um conceito "pico" (um conceito muito frequente num período de tempo) está relacionado com a diminuição da frequência de outro conceito. Já na associação direta, a existência de um conceito pico causa um aumento da frequência de outro conceito. Esse dois casos podem ser vistos na Figura **7**.

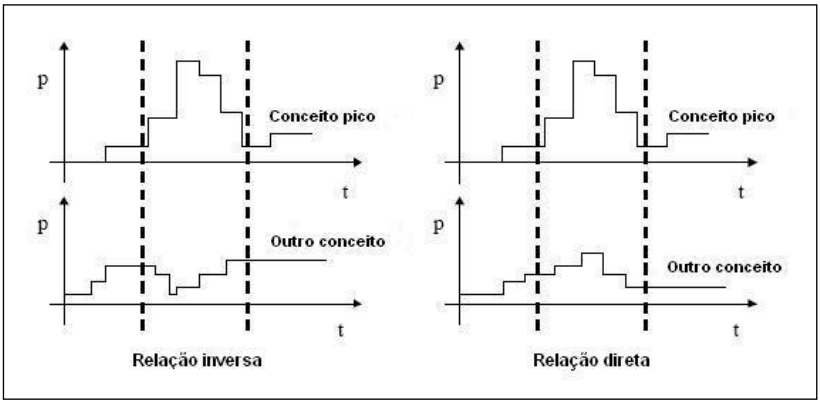


Figura 7 – *Ephemeral Association* inversa e direta.
Fonte: adaptado de (MONTES-Y-GÓMEZ; GELBUKH et al., 2001)

Esses mesmos autores também apresentam o conceito de *Deviation Detection* (MONTES-Y-GÓMEZ; GELBUKH et al., 2001). Trata-se, em MT, de um método que visa descobrir elementos irregulares em grandes quantidades de dados textuais. No caso específico de análises temporais, o objetivo é análise de situações em que há uma tendência entre dois períodos de tempo e um conceito possui um comportamento significantemente diferente desta tendência e, então, tal conceito é um "desvio" (FELDMAN; DAGAN, 1995; ARNING; RAGHAVAN, 1996; FELDMAN; AUMANN et al., 1998; KNORR; NG et al., 2000; MONTES-Y-GÓMEZ; GELBUKH et al., 2001; KAMARUDDIN; HAMDAN et al., 2007).

Há também os trabalhos que envolvem Regras de Associação Temporais (LEE; LIN et al., 2001; NØRVÅG; ERIKSEN et al., 2006; BOUANDAS; OSMANI, 2007; GHARIB; NASSAR et al., 2010). Usa-se conceitos tradicionalmente utilizados em Regras de Associação para descobrir relações temporais. Isso significa dizer que se um conceito "A" está presente em um documento no tempo $t_n$ então o conceito "B" estará presente em algum documento no tempo $t_{n+1}$. Por exemplo, a análise de registros médicos para encontrar relacionamentos entre remédios, sintomas e doenças.

Há também as abordagens visuais para AT. Uma delas chama-se *Trend Graph* (FELDMAN; AUMANN et al., 1998; SAGA; TSUJI et al., 2010). Trata-se de uma ferramenta visual que permite ao usuário ver graficamente a *evolução* e *mudanças* relacionamentos entre conceitos no

tempo. É possível comparar grafos/redes de diferentes períodos de tempo. Assim, auxilia o usuário a encontrar *tendências* e *descontinuidades* de forma visual. Outro trabalho com abordagem visual é o ThemeRiver® (HAVRE; HETZLER et al., 2002), que pode ser visto na Figura **8**. É usado para ver as mudanças temáticas ao longo do tempo em uma coleção de documentos.

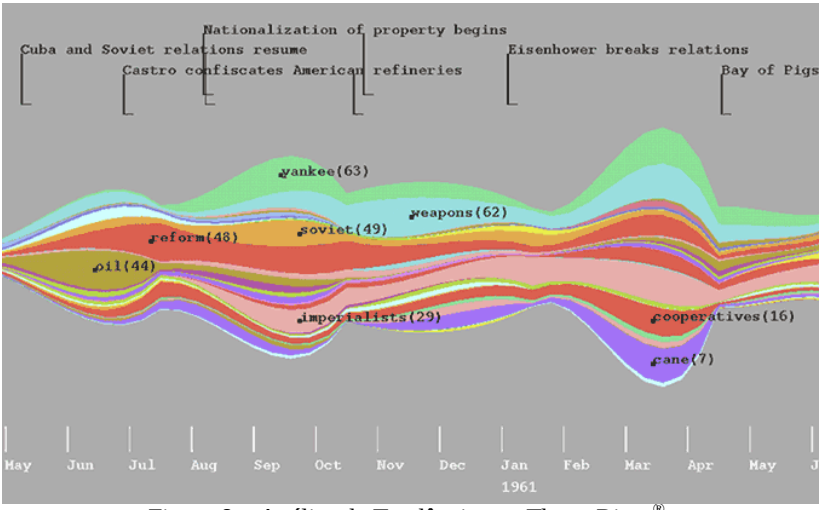


Figura 8 – Análise de Tendências no ThemeRiver®
Fonte: (HAVRE; HETZLER et al., 2002)

Uma revisão da literatura sobre abordagens visuais para informações textuais e temporais é apresentada por Šilić e Dalbelo Bašić (2010). Esse trabalho apresenta áreas relacionadas, tipos de coleções de dados que são visualizados, aspectos técnicos de geração de visualizações e metodologias de avaliação.

## 2.3 MODELOS BASEADOS EM COOCORRÊNCIA

Modelos baseados em coocorrências tem sido utilizados tradicionalmente na identificação de colocações[5] em textos (MANNING; SCHÜTZE, 1999). Esses modelos partem do pressuposto que é possível identificar estatisticamente um possível relacionamento entre palavras, analisando suas frequências individuais e conjuntas. Esse conceito pode ser generalizado para ser usado com o objetivo de se determinar a força do relacionamento entre dois elementos textuais que aparecem conjuntamente em algum documento textual (GONÇALVES, 2006). Abaixo são apresentados os principais modelos baseados em coocorrências.

### 2.3.1 Frequência

A abordagem mais simples para estabelecer a relação entre dois elementos textuais é contagem da frequência conjunta. O fato de dois elementos textuais aparecem frequentemente juntos em uma determinada coleção de documentos é uma evidência de relacionamento. O problema deste método é que ele tende a encontrar muitas combinações de palavras do tipo "of the", "in the" ou "is a". Veja a Tabela 2.

| $C(t_1,t_2)$ | $t_1$ | $t_2$ |
|---|---|---|
| 80874 | of | the |
| 58841 | in | the |
| 26430 | to | the |
| ... | | |

[5] Do inglês *collocations*. Em Linguística, uma combinação de palavras relacionadas dentro de uma sentença que ocorrem mais frequentemente do que seria possível predizer em um arranjo aleatório de palavras; uma combinação de palavras que ocorrem com frequência suficiente para serem reconhecidas como uma combinação comum, especialmente um par de palavras em que essas palavras ocorrem de maneira adjacente uma a outra (*Collaborative International Dictionary of English*, CIDE).

| | | |
|---|---|---|
| 12622 | from | the |
| 11428 | New | York |

Tabela 2 – Exemplo de frequências conjuntas extraído de uma coleção de documentos.
Fonte: (JUSTESON; KATZ, 1995)

Uma alternativa simples é a eliminação dos pares de palavras constantes em uma tabela de controle (*stop lists*). Outra forma que tende a melhorar os resultados é proposta por Justeson e Katz (1995), na qual são utilizados padrões que identificam prováveis estruturas frasais. Neste método existem três unidades que compõem os padrões: adjetivo (A), nome (N) e preposição (P). Através do uso destes padrões, os resultados melhoram consideravelmente, como pode ser observado na Tabela 3. Agora pares de palavras tais como "*New York*" possuem maior relevância.

| C($t_1$,$t_2$) | $t_1$ | $t_2$ | Padrão |
|---|---|---|---|
| 11428 | New | York | AN |
| 5412 | Los | Angeles | NN |
| 3301 | last | year | AN |
| ... | | | |

Tabela 3 – Exemplo de frequências conjuntas extraído de uma coleção de documentos.
Fonte: (JUSTESON; KATZ, 1995)

### 2.3.2 Média e Variância

Embora o uso de frequência conjunta releve indícios para a formação de estruturas frasais, muitas dessas estruturas ocorrem de maneira mais flexível, em que palavras são conectadas através de janelas. A quantidade de palavras que aparece entre outras duas palavras varia, e a distância entre elas não é a mesma. A utilização de janelas (quantidade de palavras em cada um dos lados de uma determinada palavra) oferece a solução. Como exemplo consideram-se duas palavras $t_1$ e $t_2$ que ocorrem com diferentes deslocamentos ao longo da coleção

de documentos, sendo esses deslocamentos 5, 5, 3, 4, 4, respectivamente. Nesse sentido, a média e a variância podem determinar o grau de relacionamento entre as palavras. A média é computada utilizando-se os deslocamentos, como mostrado a seguir.

Embora a frequência conjunta de duas palavras seja um indício de formação estruturas frasais, muitas dessas estruturas são formadas de maneira flexível, na qual as palavras coocorrem dentro de janelas no texto, e não de forma adjacente. A distância na qual duas palavras coocorrem no texto varia ao longo da coleção de documentos. Assim, a média e a variância da distância podem determinar o grau de relacionamento entre as palavras. Por exemplo, se duas palavras coocorrem 4 vezes em um coleção de documentos com distâncias 5, 4, 5 e 3, a média das distâncias será calculada da seguinte forma:

$$\frac{(5+4+5+3)}{4} = 4{,}25$$

E a variância informa o grau de desvio das distâncias a partir da média, sendo estimada conforme a seguinte equação:

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \quad (1)$$

onde $n$ é o número de vezes que as duas palavras coocorrem, $d_i$ é a distância da $i_{\text{th}}$ coocorrência, e $\bar{d}$ é a média das distâncias. Caso as distâncias sejam sempre as mesmas, a variância será zero. Do contrário, se as distâncias acontecem aleatoriamente, ou seja, não configuram um padrão de relacionamento, a variância será alta.

Assim, o desvio padrão $s = \sqrt{s^2}$ é utilizado para avaliar a variabilidade das distâncias entre duas palavras ou qualquer outra estrutura textual. Para o exemplo acima, o resultado é o seguinte:

$$s = \sqrt{\frac{((5-4.25)^2 + (4-4.25)^2 + (5-4.25)^2 + (6-4.25)^2)}{3}} \cong 1.3228$$

A informação provida pela média e pela variância das distâncias entre duas palavras na coleção de documentos pode ser utilizada na determinação de estruturas frasais com baixo desvio padrão. Valores de

desvios baixos indicam que duas palavras tendem a ocorrer quase sempre na mesma distância, enquanto que o valor zero indica que duas palavras ocorrem exatamente sempre na mesma distância. Esse padrão de comportamento pode indicar um relacionamento relevante entre as duas palavras. Por sua vez, valores de desvios altos indicam relacionamentos pouco relevantes.

### 2.3.3 Teste de Hipótese

Apesar de altas frequências e baixas variâncias serem indícios de relacionamentos entre palavras, não existe garantia de que isso conduza a resultados melhores dos que aqueles obtidos ao acaso. O objetivo é identificar se duas palavras ocorrem juntas mais frequentemente do que ao acaso. Avaliar se algo é ou não um evento ao acaso é um problema clássico da estatística chamado de teste de hipótese (MANNING; SCHÜTZE, 1999).

No teste de hipótese, formula-se a *hipótese nula* $H_0$ que não há uma associação entre duas palavras além das ocorrências ao acaso, calcula-se a probabilidade $p$ que o evento ocorreria se $H_0$ fosse verdadeira, e então se rejeita $H_0$ se $p$ é muito baixa (normalmente se abaixo de um nível de significância de $p < 0.05$, 0.01, 0.005, ou 0.001) e, caso contrário, se aceita $H_0$ como sendo possível. Assim, quando a hipótese nula é rejeitada, considera-se que existe um relacionamento entre as duas palavras além das ocorrências ao acaso e, de maneira similar, quando se aceita a hipótese nula considera-se que não existe um relacionamento entre as duas palavras.

### 2.3.4 Teste *t*

O teste *t* tem sido muito utilizado na identificação de colocações. Ele indica o quão provável ou improvável é a ocorrência de um determinado evento. Por meio da média e da variância, a hipótese nula é avaliada informando que a amostra é composta a partir de uma distribuição com média μ. Logo, obtém-se o resultado a partir da análise das diferenças entres as médias observadas e esperadas, normalizadas pela variância dos dados. Assim, a probabilidade da amostra para a estatística t é calculada como:

$$\boldsymbol{t} = \frac{\bar{\boldsymbol{x}} - \boldsymbol{\mu}}{\sqrt{\frac{\boldsymbol{s^2}}{\boldsymbol{N}}}} \quad (2)$$

onde $\bar{x}$ é a média da amostra, e $s^2$ é a variância da amostra, N é quantidade de pares de palavras (bigramas) existentes na coleção de documentos e $\mu$ é a média da distribuição. Se o teste *t* é grande o suficiente, a hipótese nula pode ser rejeitada. Isso significa que a relação entre os elementos textuais pode ser confirmada.

Normalmente o teste *t* é aplicado à amostra de dados. Contudo, para quando se deseja identificar colocações, existe uma forma padronizada de estendê-lo para uso de proporções e contagens. Nesse contexto, uma coleção de documentos é avaliada com um sequência de N pares de palavras. As amostras são obtidas considerando 1 (um) quando o par de interesse ocorre e 0 (zero), caso contrário.

A partir da estimativa da máxima probabilidade, é possível calcular as probabilidades de cada componente do par de palavras. Para ilustrar o seu funcionamento, tem-se o exemplo apresentado por Gonçalves (2006): tomam-se as palavras $t_1$="Inteligência" e $t_2$="Artificial" de uma determinada coleção de documentos, na qual $t_1$ ocorre 14.902 vezes e $t_2$ ocorre 6.484 vezes, em um total de 15.806.525 palavras.

$$P(t_1) = \frac{14.902}{15.806.252}$$

$$P(t_2) = \frac{6.484}{15.806.252}$$

Inicialmente a hipótese nula informa que as ocorrências de $t_1$ e $t_2$ são independentes.

$$H_0: P(t_1 t_2) = P(t_1)P(t_2) = \frac{14.902}{15.806.252} \times \frac{6.484}{15.806.252} \approx 3.8675 \times 10^{-7}$$

Assumindo que existam 32 ocorrências de "inteligência artificial" entre os 15.806.252 pares de termos da coleção de documentos, a média seria: $\bar{x} = \frac{32}{15.806.252} \approx 2.02452 \times 10^{-6}$. Utilizando esses valores na Equação 2, tem-se o seguinte valor para o teste *t*:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{P(t_1 t_2) - P(t_1)P(t_2)}{\sqrt{\frac{P(t_1 t_2)}{N}}}$$

$$\approx \frac{2.02452 \times 10^{-6} - 3.8675 \times 10^{-7}}{\sqrt{\frac{2.02452 \times 10^{-6}}{15.806.252}}} \approx 4.576208$$

O valor *t* de 4.576208 é maior que 2.576, que é valor crítico para $\alpha = 0.005$. Desse modo, a hipótese nula que "inteligência" e "artificial" ocorrem independentemente pode ser descartada. Logo, isso indica que "inteligência artificial" não é meramente composta ao acaso e possui um significado adicional, ou seja, trata-se de uma colocação.

### 2.3.5 Teste de Pearson - *Chi-square* ($\chi^2$)

O *Chi-square* ($\chi^2$) é uma técnica estatística utilizada para determinar se a distribuição das frequências observadas difere das frequências esperadas. Se a diferença entre as frequências observadas e esperadas é alta, então a hipótese nula de independência pode ser rejeitada. Isso significa que há uma relação entre os dois termos, e não apenas algo aleatório. Sua aplicação baseia-se na utilização de uma tabela 2*2 (tabela de contingência), como a apresentada na Tabela 4.

| | $w_2$ | $\overline{w}_2$ |
|---|---|---|
| $w_1$ | $a$ | $b$ |
| $\overline{w}_1$ | $c$ | $d$ |

Tabela 4 – Tabela de contingência de 2x2.

A célula *a* indica o número de vezes que $w_1$ e $w_2$ ocorrem conjuntamente, *b* indica o número de vezes que $w_1$ ocorre mas $w_2$ não, *c* é o número de vezes que $w_2$ ocorre mas $w_1$ não, e *d* é o número de documentos da coleção menos o número de vezes que nem $w_1$ e nem $w_2$ ocorrem, sendo *d=N-a-b-c*, onde *N* é o tamanho da base.

A estatística $\chi^2$ soma a diferença entre os valores observados e esperados, divididos pelos valores esperados:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{\left(a - \frac{(a+b)(a+c)}{N}\right)^2}{\frac{(a+b)(a+c)}{N}} + \frac{\left(b - \frac{(a+b)(b+d)}{N}\right)^2}{\frac{(a+b)(b+d)}{N}}$$

$$+ \frac{\left(c - \frac{(c+d)(a+c)}{N}\right)^2}{\frac{(c+d)(a+c)}{N}} + \frac{\left(d - \frac{(c+d)(b+d)}{N}\right)^2}{\frac{(c+d)(b+d)}{N}} \quad (3)$$

$$= \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Considerando a Tabela 5 como sendo a tabela de contingência que representa a distribuição para as palavras "inteligência" e "artificial":

| | $w_2$ = artificial | $\overline{w}_2 \neq$ artificial |
|---|---|---|
| $w_1$ = inteligência | **32** | $14.902 - 32 = \mathbf{14.870}$ |
| $\overline{w}_2$ = inteligência | $6.484 - 32 = \mathbf{6.452}$ | $15.806.252 - 14.870 - 6.452 - 32 = \mathbf{15.784.898}$ |

Tabela 5 – Tabela de contingência para a dependência das palavras $t_1$="inteligência" e t2="artificial".

E utilizando valores dessa tabela na Equação 3, têm-se:

$$\chi^2 = \frac{15.806.252 \times (32 \times 15.784.898 - 14.870 \times 6.452)^2}{(32 + 14.870) \times (32 + 6.452) \times (14.870 + 15.784.898) \times (6.452 + 15.784.898)} = 109.77$$

A hipótese nula indica inicialmente que as ocorrências das palavras $t_1$="inteligência" e $t_2$="artificial" são independentes. Apartir da distribuição de $\chi^2$, pode-se verificar que para o nível de probabilidade de $\alpha = 0.05$ o valor crítico de $\chi^2$ é 3.841. Como nesse exemplo o valor de $\chi^2$ está acima de 3.841, a hipotése nula pode ser rejeitada, ou seja, existe um relacionamento entre $t_1$ e $t_2$.

### 2.3.6 *Phi-squared* ($\phi^2$)

O *phi-squared* também utiliza uma tabela de contingência, similar ao método anterior. Segundo Conrad e Utt (1994), o $\phi^2$ tende a

favorecer associações com alta frequência. O *Phi-squared* (CHURCH; GALE, 1991) é definido como:

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (4)$$

onde $0 \leq \phi^2 \leq 1$.

### 2.3.7 *Informação Mútua*

O Informação Mútua (IM) tem motivação na teoria da informação e tem sido usado na identificação de relacionamentos entre palavras através de suas coocorrências na coleção de documentos (CHURCH; HANKS, 1990). O IM compara a probabilidade de um par de palavras (ou qualquer outra unidade linguística) aparecer mais frequentemente de maneira conjunta do que isoladamente. Essa medida cresce à proporção que a frequência conjunta também cresce. Se uma determinada palavra tende a ocorrer individualmente, então IM será um número negativo. A fórmula para o cálculo do IM é definida como:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N} \times \frac{f(y)}{N}} \quad (5)$$

onde *P(x,y)* é a probabilidade das palavras *x* e *y* ocorrerem conjuntamente, *P(x)* e *P(y)* são as probabilidades de *x* e *y* ocorrerem individualmente, e *N* é o tamanho da coleção de documentos. Quando existe um relacionamento forte entre as duas palavras, *I(x,y)* será maior que 0. Para exemplificar o cálculo, a máxima probabilidade é utilizada na determinação da probabilidade de dois eventos que ocorrem conjuntamente. Considere o seguinte exemplo:

$$I(Inteligência, Artificial) = \log_2 \frac{\frac{32}{15.806.252}}{\frac{14.902}{15.806.252} \times \frac{6.484}{15.806.252}} \approx 2.38$$

Do ponto de vista da teoria da informação, o IM informa que a quantidade de informação da ocorrência de “Inteligência” na posição *i*

da coleção aumenta em 2.38 bits se é aceito que “Artificial” ocorre na posição *i* + 1, ou vice-versa.

### 2.3.8 Outros modelos

Além dos modelos explicados anteriormente, existem outros modelos baseados em coocorrência. Alguns desses, são derivações do IM como, por exemplo, os trabalhos de Vechtomova et al. (2003) e Wang e Vechtomova (2005), que introduzem um parâmetro adicional: o tamanho da janela entre o par de palavras. Uma janela é definida como um número fixo de palavras à direita e à esquerda de uma determinada palavra.

Há também o método *Z score*, que promove uma indicação sobre a validade do relacionamento entre elementos textuais medindo-se a distância dos desvios padrão entre as frequências observadas das ocorrências de *y* em torno de *x* e as frequências esperadas (VECHTOMOVA; ROBERTSON et al., 2003).

Existe ainda modelos com base mais empírica, como o algoritmo CORDER (ZHU; GONÇALVES et al., 2005) e o *Latent Relation Discovery* (LRD) (GONÇALVES; ZHU et al., 2006); e o modelo de Indexação de Semântica Latente (ISL), que tem por objetivo capturar a estrutura semântica de coleções de documentos através da correlação de termos e documentos (DEERWESTER; DUMAIS et al., 1990; DING, 2000).

## 2.4 RECUPERAÇÃO DE INFORMAÇÃO

De acordo com Salton (1968), a Recuperação de Informação (RI) “é a área de pesquisa que se preocupa com a estrutura, análise, organização, armazenamento, recuperação e busca de informação”. Mitra e Chaudhuri (2000) afirma que o aumento excessivo de informações resulta em grande demanda por meios eficientes e eficazes de organização, indexação e recuperação dessa informação. A representação e organização dessa informação devem permitir que os usuários tenham acesso fácil e rápido à informação desejada. E de acordo com Kowalski (1997), o principal objetivo de um sistema de RI é minimizar a dificuldade do usuário em localizar a informação

requisitada. Segundo Baeza-Yates e Ribeiro-Neto (1999), a recuperação, representação, armazenamento, organização e acesso são os principais processos na gestão da informação. Assim, tais processos devem ser atendidos de modo a prover aos usuários a recuperação da informação almejada.

Dentro desse contexto, a RI tem como tarefa principal possibilitar a localização de documentos que satisfaçam determinada consulta efetuada pelo usuário. Para que isso seja possível, os documentos devem ter uma representação lógica que permita que as buscas sejam realizadas. Normalmente, os documentos são representados por meio de índices formados pelos termos que compõem esses documentos (RIJSBERGEN, 1979). De acordo com Baeza-Yates e Ribeiro-Neto (1999), "um índice é uma estrutura de dados crítica porque permite rápida busca sobre grandes volumes de dados". A partir do índice criado, o usuário pode descrever sua necessidade por meio de uma consulta formada por termos. O sistema de RI, então, interpreta essa consulta e a aplica sobre o índice. O resultado desse processo é uma lista de documentos que estão ordenados de acordo com algum critério predeterminado. O sistema de RI, então, apresenta essa lista de documentos ao usuários. Os detalhes do funcionamento desse processo depende do modelo de RI utilizado. Um dos modelos mais comuns é o modelo vetorial, descrito na próxima seção.

### 2.4.1 Modelo Vetorial

O Modelo Espaço Vetorial (MEV) é um dos modelos mais utilizados em aplicações de RI (MANNING; SCHÜTZE, 1999). No MEV, cada lista de termos (dos documentos ou das consultas) é considerada como um vetor de termos no espaço *n*-dimensional, onde *n* é o número de termos distintos (RUSSEL; NORVIG, 1995). O conjunto de vetores forma a matriz termo–documento que pode ser armazenada, por exemplo, na forma de um índice invertido.

Cada termo do MEV possui um peso que representa a sua relevância no documento do qual foi extraído. Entre as formas de se calcular tais pesos, a *tf-idf* (*term frequency / inverted document frequency*) é mais utilizada. Nessa forma de se calcular os pesos deve-se dividir o número de vezes que o termo aparece no documento pelo número de documentos que contém o termo. Assim, cada documento vai

ter um vetor com os seus termos e respectivos pesos. Para ser possível recuperar documentos, é também necessário criar o vetor com os termos da consulta do usuário. A partir disso, deve-se calcular a similaridade entre o vetor da consulta do usuário e os vetores dos documentos. Segundo Korfhage (1997), quando o modelo vetorial é utilizado, a medida de similaridade pode ser associada com a (a) noção de distância, por meio da qual documentos que se encontram próximos no espaço vetorial são altamente similares; ou (b) com uma medida angular, baseada na ideia de que documentos na mesma direção estão relacionados. Assim, os documentos de retorno da consulta, apresentados ao usuário, são classificados de acordo com a medida de similaridade, que representa a relevância dos documentos em relação a consulta. O modelo vetorial é considerado flexível, pois facilmente possibilita que documentos recuperados possam ser classificados e avaliados de acordo com a sua relevância (NOUALI; BLACHE, 2003).

### 2.4.2 Similaridade entre Vetores

Como discutido na seção anteriormente, o processo de cálculo da similaridade entre o vetor de termos da consulta e os vetores de termos dos documentos é essencial para se recuperar os documentos mais relevantes para o usuário. Egghe e Michel (2002) apresentam um conjunto de equações utilizadas no cálculo de similaridade, entre elas, índice Jaccard, índice Dice, medida *overlap* (máxima e mínima), medida do cosseno e medida do pseudo-cosseno. Uma discussão ampla sobre medidas de similaridade é também apresentada por Jones e Furnas (1987).

Entre tais formas de cálculo de similaridade, o cosseno tem sido muito aplicado a sistemas de RI (SALTON; BUCKLEY, 1988). A equação do cosseno mede o ângulo entre dois vetores, variando de 1.0 (cos(0°) = 1.0) para vetores apontando na mesma direção, 0.0 (cos(90°) = 0.0) para vetores ortogonais e -1.0 (cos(180°) = -1.0) para vetores apontando em direções opostas, sendo definido como:

$$\cos\theta = \frac{\sum_{i=1}^{n}(t_i \times q_i)}{\sqrt{\sum_{k=1}^{n}(t_k)^2} \times \sqrt{\sum_{j=1}^{n}(q_j)^2}} \quad (6)$$

onde $t_i$ e $t_k$ são as frequências normalizadas dos $i_{th}$ e $k_{th}$ termos do vetor $t$, e $q_i$ e $q_j$ são as frequências dos $i_{th}$ e $j_{th}$ termos do vetor $q$.

### 2.4.3 Índice Invertido

Quando se lida com grandes quantidades de documentos textuais é necessário utilizar técnicas de RI para ser possível localizar, de forma eficiente, documentos que contenham determinado termo. Segundo Baeza-Yates e Ribeiro-Neto (1999), o índice invertido é a estrutura mais comum para indexar informação de modo a permitir um bom desempenho durante uma tarefa de busca.

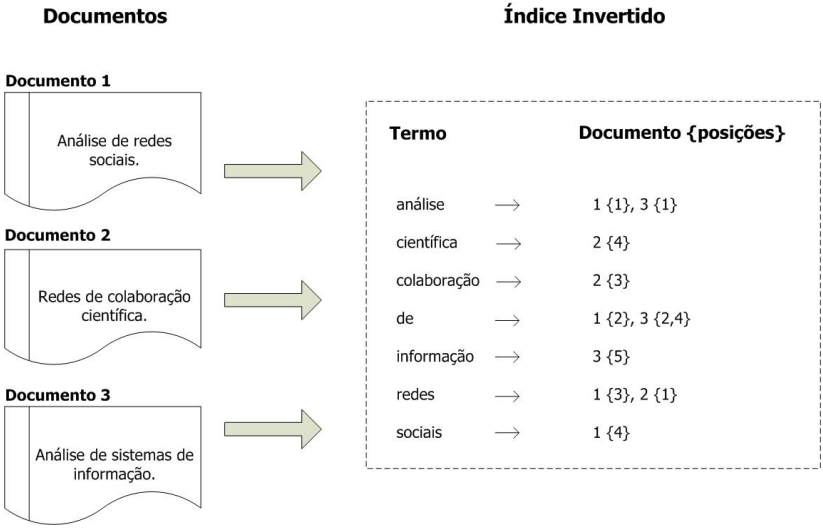


Figura 9 – Exemplo de índice invertido para três documentos.

O índice invertido possui um lista de termos presentes nos documentos textuais. Cada termo dessa lista aponta para outra lista com os respectivos documentos que o contém e suas posições dentro do documento. A Figura 9 mostra um exemplo de índice invertido que representa três documentos textuais. A partir dessa estrutura é possível, por exemplo, fazer uma busca com o termo “redes” e encontrar o documento 1 (com a palavra “redes” na posição 3) e o documento 2 (a palavra “redes” na posição 1).

## 2.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado o referencial teórico do modelo proposto. O capítulo foi dividido em três partes: descoberta de conhecimento em textos, modelos baseados em coocorrências e recuperação de informação. A primeira parte descreveu a área de descoberta de conhecimento em textos, primeiro apresentaram-se conceitos sobre KDD e MD, que estão no âmbito de dados estruturados, para, em seguida, apresentar os conceitos de KDT e MT. Também se discutiu os conceitos e trabalhos em KDT e MT que tem relação direta com os objetivos deste trabalho: correlação (relacionamentos diretos) e associação (relacionamentos indiretos) de elementos textuais; e a análise temporal de informações textuais. Na segunda parte foram apresentados em detalhes alguns dos principais modelos baseados em coocorrências que são utilizados para o cálculo da correlação e associação. Na terceira parte são apresentados conceitos da área de recuperação de informação que são utilizados neste trabalho: representação vetorial, similaridade de vetores e a estrutura de índice invertido.

# 3 MODELO PROPOSTO

Este capítulo apresenta o modelo de *Temporal Knowledge Discovery in Texts* (TKDT) proposto neste trabalho. Este modelo é iterativo e dividido por fases, assim como os modelos de KDT apresentados no Capítulo 2. O objetivo do modelo é 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. Essas tarefas são baseadas na evolução dos relacionamentos diretos e indiretos entre elementos textuais ao longo do tempo.

## 3.1 MODELO DE TKDT PROPOSTO

A Figura 10 ilustra o modelo de TKDT proposto nesta tese.

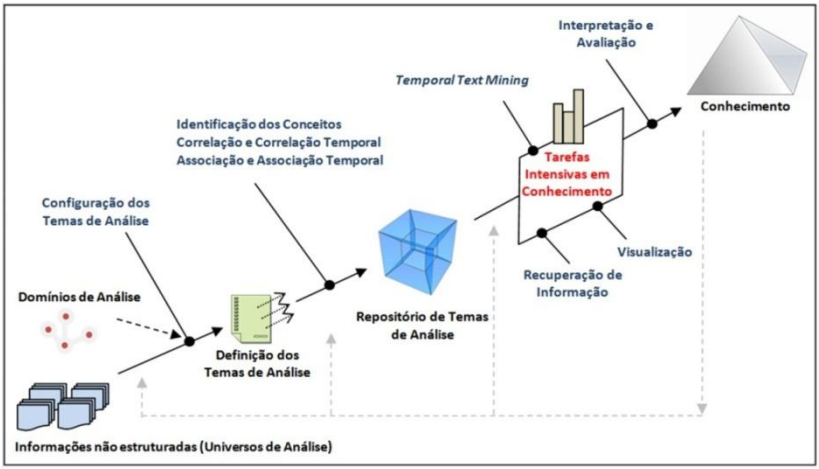


Figura 10 – Modelo de *Temporal Knowledge Discovery in Texts* proposto.

A seguir são apresentadas as diversas fases do modelo:

**Configuração dos Temas de Análise**: esta é a fase em que são configurados os temas de análise de interesse do usuário. Um tema de análise consiste em um universo de análise e um domínio de análise. O universo de análise corresponde às fontes de informação a serem

analisadas. O domínio de análise corresponde ao conhecimento de domínio utilizado, que pode estar representado por uma ontologia, tesauro, dicionário, vocabulário, etc.

**Identificação das Ocorrências dos Conceitos**: fase responsável pela identificação dos conceitos do domínio de análise nas fontes de informação (universo de análise). Consiste em localizar as ocorrências dos conceitos nos documentos textuais e na associação de uma marcação de tempo (*timestamp*) à essa ocorrência.

**Correlação e Correlação Temporal**: fase responsável pelo cálculo da força de correlação e correlação temporal entre os conceitos do domínio de análise, que foram extraídos das fontes de informação, para cada tema de análise. Para isso, um dos modelos baseados em coocorrências, apresentados no Capítulo 2 deste de trabalho, deve ser aplicado utilizando-se a frequência individual de cada conceito (número de vezes que o conceito aparece na fonte de informação), a frequência conjunta de dois conceitos (número de vezes que dois conceitos coocorrem um documento) e o tamanho do *corpus* (número de documentos na fonte de informação). O resultado final desta fase são as matrizes de correlação e correlação temporal.

**Associação e Associação Temporal**: a partir das matrizes de correlação e correlação temporal obtidas na fase anterior, é realizado o cálculo da força de associação e associação temporal de par de conceitos. Para isso, utilizam-se funções de cálculo de similaridade entre vetores, também apresentados no Capítulo 2. O resultado final desta fase são as matrizes de associação e associação temporal.

**Repositório de Temas de Análise**: o repositório de temas de análise do modelo é representado como um hipercubo de cinco dimensões. Têm-se duas dimensões de conceitos, uma para representar o conceito de origem (*source concept*) e outra o conceito destino (*target concept*); uma dimensão para representar os tipos de relacionamento; uma dimensão para representar o tempo; e uma dimensão para representar os temas de análise.

**Tarefas Intensivas em Conhecimento**: a partir do repositório de temas de análise obtido na fase anterior, várias tarefas intensivas em conhecimento, com ênfase em relacionamentos temporais entre os conceitos, podem ser realizadas. A definição das tarefas, suas ferramentas, métodos e algoritmos são baseados na literatura sobre correlação, associação e análise temporal, apresentada no Capítulo 2. Essa fase envolve a participação dos usuários na interpretação e

avaliação dos resultados. Nas seções seguintes estas fases são explicadas em detalhes.

## 3.2 CONFIGURAÇÃO DOS TEMAS DE ANÁLISE

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. Um documento pode ser definido como uma unidade discreta de dados textuais que normalmente, mas não necessariamente, está relacionado a um documento do mundo real. Um documento pode ser, por exemplo, um e-mail, um relatório ou artigo científico.

A seguir são apresentados alguns exemplos de informações não estruturadas que podem ser utilizadas:

- **Mensagens de e-mails**. As mensagens enviadas e recebidas pelo endereço de e-mail corporativo.
- **Mensagens instantâneas**. Históricos de mensagens trocadas entre colaboradores através de softwares de mensagens instantâneas.
- **Registros de buscas**. Os termos buscados no sistema de busca textual da instituição ou mesmo em motores de busca (search engines), como o Google®, podem conter informações úteis sobre os interesses e necessidades dos colaboradores de tal instituição.
- ***World Wide Web***. O conteúdo de páginas Web como, por exemplo, sites de notícias, páginas pessoais, blogs, wikis, sítios governamentais, etc.
- **Artigos científicos**. Bases de artigos científicos nas mais diversas áreas, como por exemplo, Scielo6 e Medline7.
- **Campos textuais em bancos de dados estruturados**. Muitos bancos de dados estruturados contêm campos com informações não-estruturadas ou semiestruturadas. Por exemplo, uma tabela

6 http://www.scielo.org.
7 http://www.ncbi.nlm.nih.gov/pubmed/

produto com um campo chamado descricao que contém uma descrição em linguagem natural sobre o produto.

- **Documentos eletrônicos em geral**. Manuais, relatórios técnicos, projetos, currículos, e-books, etc.

Já o domínio de análise refere-se ao conhecimento de domínio utilizado nas análises. O domínio de análise é formado por um conjunto de instâncias da área de interesse. Uma área de interesse pode ser uma especialidade do conhecimento (ex.: Medicina), um setor de uma organização (ex.: Recursos Humanos) ou qualquer domínio que tenha um contexto, uma semântica e um conjunto de informações disponíveis. O conhecimento de domínio é parte importante do modelo, pois reduz o espaço de buscas, uma vez que somente os relacionamentos entre os conceitos pertencentes ao domínio serão recuperados das fontes de informação. Esse conhecimento de domínio pode estar representado em ontologias, tesauros, taxonomias, dicionário, vocabulários, etc. O modelo permite que se criem diversos temas de análise a partir da combinação entre diferentes universos e domínios de análise. Cada tema representa uma visão diferente que o usuário terá sobre as fontes de informação, de acordo com cada domínio de análise escolhido. Isso permite a geração de análises flexíveis dependendo dos interesses do usuário.

### 3.2.1 Exemplo de Tema de Análise

Uma coleção de resumos de artigos relacionados à área de Ciência da Informação, com data de publicação entre o ano de 2005 e 2008, pode ser considerada como um universo de análise, pois se trata de uma de uma coleção de documentos textuais com um atributo temporal. Já o domínio de análise pode ser representado por um conjunto de instâncias da ontologia mostrada na Figura 11. Essa ontologia possui cinco classes: *Keyword*, *Paper*, *Author*, *Institution* e *Journal*.

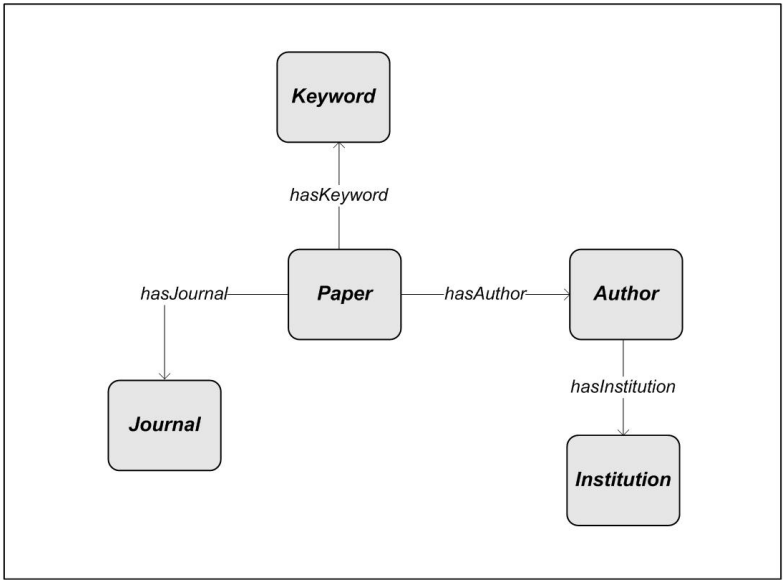


Figura 11 – Ontologia utilizada para descrever o domínio de análise.

Nesse exemplo, as instâncias que compõem o domínio de análise pertencem à classe *Keyword* e estão listadas na Figura 12, em linguagem OWL[8] (*Web Ontology Language*).

8 http://www.w3.org/TR/owl-features/

```
...
<owl:Ontology rdf:about="" />
<owl:Class rdf:ID="Keyword" />
<Keyword rdf:ID="Ciência" />
<Keyword rdf:ID="Redes" />
<Keyword rdf:ID="Conhecimento" />
<Keyword rdf:ID="Informação" />
<Keyword rdf:ID="Inovação" />
<Keyword rdf:ID="Gestão" />
<Keyword rdf:ID="Tecnologia" />
<Keyword rdf:ID="Sistema" />
<Keyword rdf:ID="Metodologia" />
<Keyword rdf:ID="Qualidade" />
...
```

Figura 12 – Instâncias da classe *Keyword* representando os conceitos do domínio de análise.

Depois da definição do tema de análise, o próximo passo refere-se à identificação das ocorrências dos conceitos do domínio de análise nas fontes de informação do universo de análise.

## 3.3 IDENTIFICAÇÃO DAS OCORRÊNCIAS DOS CONCEITOS

Depois da definição do tema de análise, os próximos passos consistem na extração dos conceitos (domínio de análise) a partir das fontes de informação (universo de análise), e na identificação/associação de um *timestamp* à essa ocorrência.

A primeira parte consiste em localizar as ocorrências dos conceitos nos documentos textuais utilizando métodos da área de EI. Para exemplificar a fase de extração de conceitos, a Figura 13 apresenta um dos documentos do universo de análise citado na fase anterior (resumos de artigos da área de Ciência da Informação). Os conceitos do domínio de análise (Figura 12) encontrados foram destacados, com sua posição no texto.

A análise de redes$_{4}$ de colaboração científica sob as novas tecnologias$_{11}$ de informação$_{13}$ e comunicação: um estudo na Plataforma Lattes

As redes$_{22}$ de pesquisa impulsionam a criação do conhecimento$_{29}$ e o processo de inovação$_{34}$ resultantes do intercâmbio de informações$_{39}$ e, sobretudo, da junção de competências de grupos que unem esforços na busca de metas comuns. Este artigo apresenta um breve histórico dos estudos relativos às redes$_{66}$ de colaboração científica, sua evolução cronológica e as principais abordagens de estudo. Discute-se particularmente como as análises de redes$_{85}$ de pesquisa podem ser revisitadas à luz das possibilidades recentes surgidas com as novas Tecnologias$_{100}$ da Informação$_{102}$ e da Comunicação (TICs). Para tal, apresentam-se exemplos de sistemas de conhecimento$_{114}$ no âmbito da Plataforma Lattes: Egressos, Colaboradores e Redes-GP$_{123}$. Esses sistemas permitem executar, com grandes volumes de dados, análises de redes$_{135}$ por meio de algoritmos descritos na literatura, bem como criar novas formas de análise possibilitadas pelas TICs.

Colaboração científica; Análise de redes$_{157}$ sociais; Redes$_{159}$ de pesquisa; Tecnologias$_{162}$ da informação$_{164}$ e da comunicação.

Figura 13 – Artigo: A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes (2005)[9].

É possível saber, por exemplo, que o conceito "Redes" do domínio de análise ocorre no documento mostrado na Figura 13 nas posições "4, 22, 66, 85, 123, 135, 157, 159". Esse processo deve ser realizado para todos os conceitos do domínio de análise sobre todos os documentos do universo de análise.

A segunda parte consiste na associação de um *timestamp* às ocorrências identificadas em cada documento. Pode-se atribuir à cada ocorrência de um conceito a data do próprio documento. A data do documento pode ser obtida, por exemplo:

- Através de metadados que informem explicitamente a data do documento;

9 http://revista.ibict.br/index.php/ciinf/article/view/619/552

- Por meio de técnicas de reconhecimento de entidades (NER – *Named Entity Recognition*) para achar a data no conteúdo do próprio documento (ex.: data em cabeçalho de e-mail);
- Associando uma data ao documento através de snapshots. Em fontes de informação muito dinâmicas como, por exemplo, sites de notícias, pode-se coletar informações em períodos regulares de tempo e associar a data da coleta aos documentos. Por exemplo, as páginas de notícias podem ser coletadas diariamente e, assim, tem-se um conjunto de documentos para cada dia.

Eventualmente, também é possível obter o *timestamp* das ocorrências de cada conceito através de técnicas que possibilitem a extração de expressões temporais dos documentos (ALONSO; GERTZ et al., 2009). Cada expressão temporal deve ser normalizada e o *timestamp* obtido pode ser associado aos conceitos que coocorrem com a expressão (STROTGEN; GERTZ et al., 2010). A forma e a possibilidade de se obter uma marca temporal para cada documento dependem das características da cada fonte de informação.

As informações obtidas nesta fase informam *onde*, *quando* e *quais* conceitos do domínio de análise foram encontrados no universo de análise. Essas informações são utilizados nos processos de correlação e correlação temporal do tema de análise, que a fase a seguir.

## 3.4 CORRELAÇÃO E CORRELAÇÃO TEMPORAL

A *correlação* é entendida como o peso (força) da relação entre dois conceitos baseado em suas coocorrências nos documentos textuais da fonte de informação. Dois conceitos coocorrem quando aparecem juntos, por exemplo, nas seguintes situações:

- no mesmo documento;
- na mesma janela de tamanho n dentro do documento; ou
- em qualquer outra unidade discreta de texto dentro de uma fonte de informação como, por exemplo, parágrafos e sentenças.

Para simplificar, utilizar-se-á a expressão “coocorrência por documento” de forma genérica, sendo que “documento” pode significar qualquer uma das noções de coocorrência citadas acima.

A partir das posições dos conceitos em cada documento – obtidas na fase anterior – e considerando as noções da coocorrência apresentadas acima; é possível calcular as entradas necessárias para os modelos de correlação apresentados no Capítulo 2. Tais entradas incluem a frequência individual de cada conceito (número de vezes que o conceito aparece na fonte de informação), a frequência conjunta de dois conceitos (número de vezes que dois conceitos coocorrem um documento), e o tamanho do *corpus* (número de documentos nas fonte de informação)[10]. Além disso, as frequências individuais e conjuntas, e o tamanho do *corpus* devem ser calculados por tempo, de acordo com o *timestamp* de cada ocorrência, também obtido na fase anterior. A Tabela 6 descreve todas essa informações.

| **Informação** | **Descrição** |
|---|---|
| Frequência | Número de vezes que cada conceito foi encontrado na fonte de informação. |
| Frequência Conjunta | O número de vezes que cada possível par de conceitos apareceram juntos (coocorrência) em um mesmo documento. |
| Tamanho do *corpus* | Número total de documentos existentes na fonte de informação. |
| Frequência por Tempo | Número de vezes que cada entidade foi encontrada na fonte de informação por tempo. |
| Frequência Conjunta por Tempo | O número de vezes que cada possível par de conceitos apareceram juntos (coocorrência) em um mesmo documento por tempo. |
| Tamanho do *corpus* por Tempo | Número de documentos da fonte de informação por tempo. |

Tabela 6 – Informações necessárias para o cálculo das matrizes de correlação e correlação temporal para um tema de análise.

10 O modelo LRD (*Latent Relation Discovery*) utiliza ainda a distância entre os dois conceitos no documento textual (GONÇALVES; ZHU et al., 2006).

A seguir têm-se um exemplo de cálculo das matrizes de correlação e correlação temporal para o tema de análise utilizado como exemplo. Para este tema de análise tem-se uma fonte de informação com 86 documentos divididos em quatro anos e um domínio de análise com dez conceitos (Figura 12). Como modelo de correlação utilizou-se o *phi-squared* ($\phi^2$). Esse método necessita como entrada as frequências individuais, conjuntas e o tamanho do *corpus*.

Assim, têm-se as informações apresentadas nas Tabela 7 e Tabela 8. Na Tabela 7 as informações não consideram a dimensão tempo.

| **Conceito e a frequência** | | **Par de conceitos e a frequência conjunta** | |
|---|---|---|---|
| | | Ciência-Informação | 82 |
| Informação | 350 | Informação-Tecnologia | 64 |
| Conhecimento | 100 | Conhecimento-Informação | 54 |
| Ciência | 92 | Informação-Redes | 39 |
| Tecnologia | 67 | Conhecimento-Gestão | 37 |
| Gestão | 66 | ... | |
| Redes | 55 | ... | |
| Sistema | 42 | Qualidade-Tecnologia | 2 |
| Metodologia | 26 | Gestão-Inovação | 1 |
| Inovação | 17 | Metodologia-Qualidade | 1 |
| Qualidade | 12 | Qualidade-Redes | 1 |
| | | Qualidade-Sistema | 1 |

Tabela 7 – Frequências individuais e conjuntas.

Já na Tabela 8 os dados estão divididos por ano (data de publicação do documento).

| Ano | Conceito (frequência) | | Pares de conceitos (freq. conjunta) | |
|---|---|---|---|---|
| 2005 | Informação | 76 | Informação-Tecnologia | 13 |
| | Redes | 17 | Informação-Inovação | 12 |
| | Tecnologia | 14 | Informação-Redes | 12 |
| | Conhecimento | 13 | Conhecimento-Informação | 11 |
| | Inovação | 13 | Informação-Sistema | 11 |
| | Ciência | 11 | ... | |
| | Sistema | 11 | ... | |
| | Metodologia | 6 | Conhecimento-Metodologia | 1 |
| | Gestão | 4 | Conhecimento-Qualidade | 1 |
| | Qualidade | 1 | Inovação-Sistema | 1 |
| | | | Metodologia-Redes | 1 |
| | | | Metodologia-Tecnologia | 1 |
| 2006 | Informação | 118 | Ciência-Informação | 23 |
| | Ciência | 29 | Informação-Tecnologia | 17 |
| | Conhecimento | 25 | Conhecimento-Informação | 15 |
| | Redes | 18 | Informação-Redes | 11 |
| | Tecnologia | 17 | Gestão-Informação | 9 |
| | Sistema | 15 | ... | |
| | Gestão | 9 | ... | |
| | Metodologia | 8 | Metodologia-Sistema | 1 |
| | Qualidade | 8 | Metodologia-Tecnologia | 1 |
| | Inovação | 1 | Qualidade-Redes | 1 |
| | | | Qualidade-Sistema | 1 |
| | | | Qualidade-Tecnologia | 1 |
| 2007 | Informação | 77 | Ciência-Informação | 22 |
| | Conhecimento | 40 | Conhecimento-Gestão | 20 |
| | Ciência | 23 | Informação-Tecnologia | 18 |
| | Gestão | 22 | Informação-Sistema | 15 |
| | Tecnologia | 20 | Conhecimento-Informação | 13 |
| | Sistema | 16 | .. | |
| | Redes | 9 | .. | |
| | Metodologia | 4 | Metodologia-Sistema | 2 |
| | | | Metodologia-Tecnologia | 2 |
| | | | Ciência-Gestão | 1 |
| | | | Gestão-Metodologia | 1 |
| | | | Redes-Sistema | 1 |

| | | | | |
|---|---|---|---|---|
| 2008 | Informação | 79 | Ciência-Informação | 28 |
| | Gestão | 31 | Gestão-Informação | 16 |
| | Ciência | 29 | Informação-Tecnologia | 16 |
| | Conhecimento | 22 | Conhecimento-Informação | 15 |
| | Tecnologia | 16 | Conhecimento-Gestão | 13 |
| | Redes | 11 | ... | |
| | Metodologia | 8 | ... | |
| | Inovação | 3 | Gestão-Metodologia | 1 |
| | Qualidade | 3 | Informação-Qualidade | 1 |
| | | | Inovação-Tecnologia | 1 |
| | | | Metodologia-Qualidade | 1 |
| | | | Qualidade-Tecnologia | 1 |

Tabela 8 – Frequências individuais e conjuntas por ano.

A partir destas informações, podem-se calcular as matrizes de correlação e correlação temporal e, a partir destas, as matrizes de associação e associação temporal.

A correlação mostra a força do relacionamento direto entre dois conceitos quaisquer. Uma matriz de correlação possui tamanho *n* x *n*, onde *n* é o número de conceitos do domínio. Cada célula $w_{ij}$ dessa matriz representa a força do relacionamento entre dois conceitos (*i* e *j*), calculada a partir das suas frequências individuais, frequências conjuntas e, dependendo do modelo de correlação, através das distâncias entre os conceitos nos documentos. Veja a Figura 14.

$$\begin{matrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{matrix}$$

Figura 14 – Matriz de correlação para *n* conceitos.

Para o cálculo da matriz de correlação (Figura 14), são utilizados modelos baseados em coocorrência, como os apresentados no capítulo anterior. Para exemplificar esse cálculo, utilizou-se o método *phi-squared* ($\Phi^2$) para o cálculo da correlação entre os conceitos "Ciência" e "Informação", utilizando os dados apresentados na Tabela 7. Nessa tabela tem-se que o conceito "Ciência" tem frequência 92 e o conceito "Informação" frequência 350; e a frequência conjunta é 82. O tamanho do *corpus*, 819, é o número de documentos existentes na fonte de

informação. Assim, a partir de tais valores calcula-se a tabela de contingência, como mostrado na Tabela 9.

| | $w_2 =$ informação | $\overline{w}_2 \neq$ informação |
|---|---|---|
| $w_1 =$ ciência | $\mathbf{82}$ | $92 - 82 = \mathbf{10}$ |
| $\overline{w}_1 \neq$ ciência | $350 - 82 = \mathbf{268}$ | $819 - 82 - 10 - 268 = \mathbf{459}$ |

Tabela 9 – Exemplo de tabela de contingência para a dependência dos conceitos "Ciência" e "Informação".

E utilizando-se os valores da tabela de contingência com a Equação 4, tem-se:

$$\phi^2 = \frac{(82\times459-10\times268)^2}{(82+10)\times(268+459)\times(82+268)\times(10+459)} = \mathbf{0,111309}$$

Assim, a força de correlação entre os conceitos "Ciência" e "Informação" − de acordo com suas frequências individuais e conjuntas e utilizando o método de correlação *Phi-Squared* ($\phi^2$) − é 0,111309.

Repetindo esse cálculo para todos os pares de conceitos da Tabela 7, tem-se a matriz de correlação apresentada na Figura 15. Os conceitos estão nas linhas e colunas, representados por $C_i$ e $C_j$, sendo que *i* e *j* são inteiros numerados de acordo com a ordem na qual os conceitos aparecem na Figura 12: Ciência ($C_1$), Redes ($C_2$), Conhecimento ($C_3$), Informação ($C_4$), Inovação ($C_5$), Gestão ($C_6$), Tecnologia ($C_7$), Sistema ($C_8$), Metodologia ($C_9$) e Qualidade ($C_{10}$).

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $C_1$ | 1 | 0,027950 | 0,019312 | 0,111309 | 0,000006 | 0,001176 | 0,017865 | 0,003310 | 0,040097 | 0,002826 |
| $C_2$ | | 1 | 0,066328 | 0,023350 | 0,040183 | 0,004089 | 0,028597 | 0,025213 | 0,014011 | 0,000062 |
| $C_3$ | | | 1 | 0,007212 | 0,024015 | 0,157231 | 0,035360 | 0,006785 | 0,010536 | 0,000276 |
| $C_4$ | | | | 1 | 0,022868 | 0,004996 | 0,101448 | 0,036406 | 0,012334 | 0,003480 |
| $C_5$ | | | | | 1 | 0,000135 | 0,072374 | 0,001919 | 0 | 0 |
| $C_6$ | | | | | | 1 | 0,036050 | 0,000793 | 0,000536 | 0,005760 |
| $C_7$ | | | | | | | 1 | 0,008498 | 0,002265 | 0,001426 |
| $C_8$ | | | | | | | | 1 | 0,032012 | 0,000314 |
| $C_9$ | | | | | | | | | 1 | 0,001287 |
| $C_{10}$ | | | | | | | | | | 1 |

Figura 15 – Matriz de correlação com 10 conceitos.

A matriz de correlação apresentada na Figura 15 não considera a dimensão tempo. Nessa matriz, todos os relacionamentos são considerados como se ocorressem no mesmo ao tempo. Para se obter os relacionamentos ao longo do tempo é necessário gerar a matriz de

correlação temporal. Trata-se uma matriz que representa a força do relacionamento entre dois conceitos por tempo. Assim, tem-se uma matriz de tamanho *n* x *n* x *t*, onde *n* é o número de conceitos do domínio *t* é a dimensão tempo. Assim, cada célula $w_{ijk}$ dessa matriz representa a força de correlação entre dois conceitos (*i* e *j*) em um determinado tempo (*k*), calculada a partir das suas frequências individuais, frequências conjuntas naquele tempo. Veja a Figura 16.

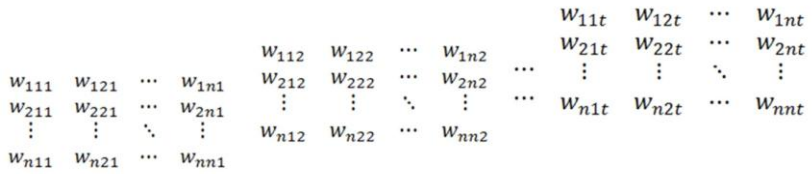

Figura 16 – Matriz de correlação temporal (*n* conceitos e *t* tempos).

O princípio é o mesmo para o cálculo da matriz de correlação temporal. A diferença é que se usam as frequências individuais e conjuntas por tempo. Assim, a partir dos dados da Tabela 8 tem-se a matriz apresentada na Figura 17. Essa matriz apresenta a correlação temporal para os 10 conceitos apresentados na Figura 12 nos quatro anos (de 2005 a 2008) disponíveis na fonte de informação. A maioria dos dados foi omitida para facilitar o seu entendimento.

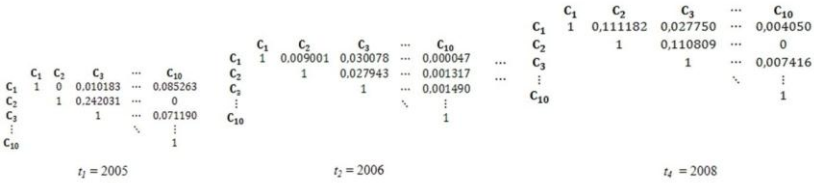


Figura 17 – Matriz de correlação temporal com 10 conceitos e 4 tempos.

## 3.5 ASSOCIAÇÃO E ASSOCIAÇÃO TEMPORAL

A *associação* é entendida como o peso (força) da relação entre dois conceitos baseado na comparação (*matching*) de seus vetores de contexto. A associação mostra a força do relacionamento indireto entre

dois conceitos quaisquer. Essa matriz tem tamanho *n* x *n*, onde *n* é o número de conceitos do domínio. Cada célula $w_{ij}$ dessa matriz representa a força do relacionamento indireto entre dois conceitos (*i* e *j*), calculada a partir da similaridade[11] entre os vetores dos dois conceitos. A matriz de associação é semelhante à matriz de correlação. A diferença está no cálculo dos pesos das relações. O valor de cada célula $w_{ij}$ de uma matriz de associação é obtido a partir do cálculo da similaridade entre os vetores de dois conceitos (*i* e *j*). O vetor de contexto de um conceito é formado pelos conceitos com os quais ele coocorre e seus respectivos pesos. Esse vetor é obtido a partir da matriz de correlação, apresentada no passo anterior. Assim, o vetor de contexto do conceito na linha *i* da matriz de correlação vai ser representado por um vetor da seguinte forma: $[w_{i1}, w_{i2},..., w_{in}]$. Por exemplo, analisando a matriz de correlação Figura 15 tem-se o vetor de contexto do conceito "Ciência", apresentado na Figura 18.

| Ciência | 1 |
|---|---|
| Informação | 0,111309 |
| Metodologia | 0,040097 |
| Redes | 0,027950 |
| Conhecimento | 0,019312 |
| Tecnologia | 0,017865 |
| Sistema | 0,003310 |
| Qualidade | 0,002826 |
| Gestão | 0,001176 |
| Inovação | 0,000006 |

Figura 18 – Vetor de contexto de "Ciência".

A partir da representação vetorial de cada conceito – obtida a partir da matriz de correlação – calcula-se a matriz de associação. Por exemplo, para se calcular a associação entre os conceitos "Ciência" ($C_1$) e "Informação" ($C_4$), pode-se calcular a similaridade entre seus respectivos vetores (obtidos a partir da matriz de correlação da Figura 15) utilizando a função cosseno (Figura 19).

11 Similaridade entre vetores foi tratada no Capítulo 2.

| **Ciência** | | | | **Informação** | | **Cosseno** |
|---|---|---|---|---|---|---|
| ↳ | Ciência | 1 | | ↳ Informação | 1 | |
| | Informação | 0,111309 | | Ciência | 0,111309 | |
| | Metodologia | 0,040097 | | Tecnologia | 0,101448 | |
| | Redes | 0,027950 | | Sistema | 0,036406 | |
| | Conhecimento | 0,019312 | **X** | Redes | 0,023350 | **0,22134** |
| | Tecnologia | 0,017865 | | Inovação | 0,022868 | |
| | Sistema | 0,003310 | | Metodologia | 0,012334 | |
| | Qualidade | 0,002826 | | Conhecimento | 0,007212 | |
| | Gestão | 0,001176 | | Gestão | 0,004996 | |
| | Inovação | 0,000006 | | Qualidade | 0,003480 | |

Figura 19 – A similaridade entre os vetores de contexto dos conceitos "Ciência" e "Informação" calculada pela equação cosseno (Equação 6).

Logo, o peso de associação entre os conceitos "Ciência" e "Informação" é 0.22134. Realizando esse processo para todos os pares de conceitos obtém-se a matriz de associação, que é apresentada na Figura 20.

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $C_1$ | 1 | 0,027950 | 0,019312 | 0,221340 | 0,000006 | 0,001176 | 0,017865 | 0,003310 | 0,040097 | 0,002826 |
| $C_2$ | | 1 | 0,133602 | 0,054292 | 0,083900 | 0,019478 | 0,064793 | 0,052274 | 0,030816 | 0,000391 |
| $C_3$ | | | 1 | 0,022775 | 0,052423 | 0,307231 | 0,079172 | 0,016100 | 0,022852 | 0,001582 |
| $C_4$ | | | | 1 | 0,053376 | 0,014690 | 0,203448 | 0,074075 | 0,030506 | 0,007386 |
| $C_5$ | | | | | 1 | 0,006820 | 0,147087 | 0,006426 | 0,001316 | 0,000192 |
| $C_6$ | | | | | | 1 | 0,076593 | 0,003220 | 0,002966 | 0,011488 |
| $C_7$ | | | | | | | 1 | 0,021707 | 0,007481 | 0,003446 |
| $C_8$ | | | | | | | | 1 | 0,064848 | 0,000824 |
| $C_9$ | | | | | | | | | 1 | 0,002747 |
| $C_{10}$ | | | | | | | | | | 1 |

Figura 20 – Matriz de associação com 10 conceitos.

Além da matriz de associação, também deve ser calculada a matriz de associação temporal. Tal matriz representa a força do relacionamento indireto entre dois conceitos por tempo. Assim, tem-se uma matriz de tamanho *n* x *n* x *t*, onde *n* é o número de conceitos do domínio e *t* é a dimensão tempo. Assim, cada célula $w_{ijk}$ dessa matriz representa a força do relacionamento de indireto entre dois conceitos (*i* e *j*) em um determinado tempo (*k*), calculada a partir da similaridade entre os vetores de contexto dos dois conceitos naquele tempo. O princípio é mesmo para o cálculo da matriz de associação temporal. A diferença é

que se usa, neste caso, a matriz de correlação temporal. Assim, a partir da matriz de correlação temporal apresentada na Figura 17 é possível calcular a matriz de associação temporal, ilustrada parcialmente na Figura 21.

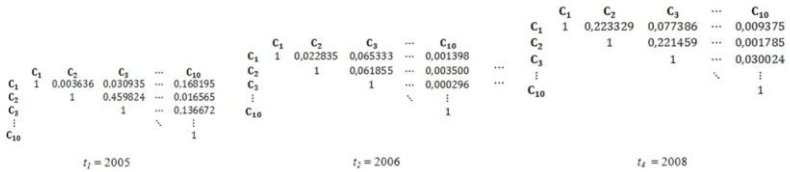


Figura 21 – Matriz de associação temporal com 10 conceitos e 4 tempos.

As matrizes de correlação temporal e associação temporal obtidas até agora possuem três dimensões: *n* x *n* x *t*. Para o modelo ter mais flexibilidade, adiciona-se mais uma dimensão: o tipo de relacionamento entre os conceitos. Assim, tem-se *n* x *n* x *t* x *r*, onde *r* significa *relation*. Essa dimensão pode ser utilizada para representar, por exemplo:

- As diferentes formas de coocorrências entre conceitos, tais como: coocorrência por documento, por janelas de diferentes tamanhos, por sentença, etc.;
- Relacionamentos que estejam definidos em uma ontologia do domínio de análise. Por exemplo, o relacionamento entre conceitos da classe *Paper* e da classe *Keyword* é, segundo a ontologia ilustrada na Figura 11, do tipo *hasKeyword*;

## 3.6 CRIAÇÃO DO REPOSITÓRIO DE TEMAS DE ANÁLISE

Todo esse processo para se obter as matrizes de tamanho *n* x *n* x *r* x *t* deve ser realizado para cada tema de análise. Portanto, obtém-se uma matriz de tamanho *n* x *n* x *r* x *t* x *a*, na qual *a* representa a dimensão tema de análise. 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*).

A Tabela 10 apresenta as cinco dimensões que definem o repositório de temas de análise.

| Dimensão | Descrição |
|---|---|
| Conceito (*source*) | Todos os conceitos do domínio de análise que foram extraídos da fonte de informação. |
| Conceito (*target*) | Todos os conceitos do domínio de análise que foram extraídos das fontes de informação. |
| Tempo | Os tempos que estão associados às ocorrências dos conceitos nas fontes de informação. |
| Relação | Os tipos de relação entre os conceitos. |
| Tema | Os temas de análise definidos pelos usuários. |

Tabela 10 – Dimensões do Repositório de Temas de Análise.

A Figura 22 apresenta uma ontologia que representa conceitualmente o repositório de temas de análise do modelo.

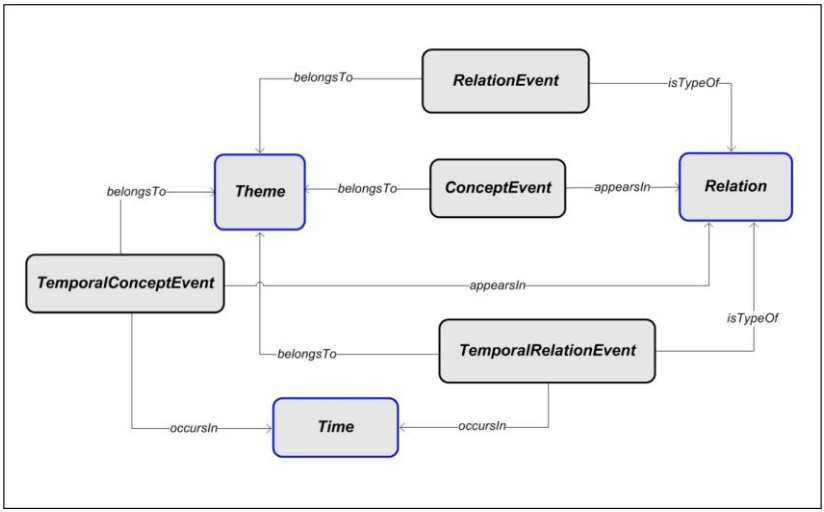


Figura 22 – Ontologia que representa o repositório de temas análise do modelo.

As classes, e suas propriedades, da ontologia apresentada na Figura 22 são descritas a seguir:

- ***Theme***. Utilizada para representar os temas de análise definidos utilizados no modelo. Possui as propriedades *id*, que identifica cada tema; e *name*, nome do tema;
- ***Relation***. Representa os tipos de relações entre dois conceitos. Possui as propriedades *id* e *name*. Possui pelo menos uma instância, chamada de *General*, que é utilizada quando o tipo da relação entre dois conceitos não é conhecido ou não é considerado.
- ***Time***. Representa a dimensão tempo em diversas granularidades (dia, mês, ano, etc.). As instâncias dessa classe estão associadas às ocorrências temporais dos conceitos e seus relacionamentos. Possui as propriedades *id* e *timestamp*.
- ***ConceptEvent***. Representa a ocorrência individual de um determinado conceito (*concept*_id) em uma determinada relação (*appearsIn*) e em um determinado tema (*belongsTo*). A frequência na qual o conceito ocorre é representado pela propriedade *frequency*.
- ***TemporalConceptEvent***. Representa a ocorrência individual de um determinado conceito (*concept*_id) em uma determinada relação (*appearsIn*), em determinado tempo (*occursIn*), e em um determinado tema (*belongsTo*). A frequência é representada pela propriedade *frequency*.
- ***RelationEvent***. Representa a ocorrência conjunta de dois conceitos (source_*concept*_id e target_*concept_id*) em um determinado tipo de relação (*isTypeOf*) e em um determinado tema (*belongsTo*). Além da frequência conjunta (*joint_frequency*), contêm a correlação (*correlation_weight*) e a associação (*association_weight*) entre os dois conceitos.
- ***TemporalRelationEvent***. Representa a ocorrência conjunta de dois conceitos (*source_concept_id* e *target_concept_id*) em um determinado tipo de relação (*isTypeOf*), em determinado tempo (*occursIn*), e em um determinado tema (*belongsTo*). Além da frequência conjunta (*joint_frequency*), contêm a correlação (*correlation_weight*) e a associação (*association_weight*) entre os dois conceitos.

## 3.7 TAREFAS INTENSIVAS EM CONHECIMENTO

O objetivo desta fase é a execução de tarefas intensivas em conhecimento com ênfase nos aspectos temporais dos relacionamentos diretos e indiretos entre os conceitos do domínio. A fundamentação teórica deste trabalho apresentou alguns métodos e técnicas das áreas de correlação, associação e análise temporal de informações textuais. Neste modelo, essas áreas são agrupadas em torno da área de *Temporal Text Mining* (TTM), que pode ser definida como a descoberta de padrões temporais em informações textuais coletadas ao longo do tempo (MEI; ZHAI, 2005). Assim, os métodos de TTM e de visualização de informações temporais, auxiliados pela área de RI, são combinados para apoiar os usuários em diversas tarefas intensivas em conhecimento. A seguir algumas dessas possíveis tarefas são apresentadas:

**Geração de Vetores de Contexto**. Consiste em se obter a lista dos conceitos mais fortemente relacionados a um dado conceito. Representa o contexto no qual o conceito ocorre na fonte de informação. Esse contexto pode ser dividido por tempo, nesse caso tem-se vetores temporais de contexto.

**Descoberta ABC**: consiste em descobrir relacionamentos entre conceitos, que apesar de não coocorrerem, estão conectados indiretamente por outros conceitos. É dividida em descoberta aberta e fechada.

**Rastreamento de Tópicos**: tarefa que consiste em detectar um tópico e rastreá-lo no tempo. Um tópico pode ser definido por um conjunto de conceitos que estão relacionado a algum assunto específico.

**Análise de Relacionamentos Temporários**: trata-se de um tipo de análise que se permite ver a influência (direta ou inversa) dos conceitos mais frequentes em um período sobre outros conceitos no mesmo período.

**Detecção de Desvios**: tarefa que visa descobrir elementos irregulares em grandes quantidades de dados textuais. No caso específico de análises temporais, concentra-se na análise de situações em que há uma tendência entre dois períodos de tempo e existe um conceito que possui um comportamento significantemente diferente desta tendência. Logo, tal conceito é considerado um "desvio".

**Extração de Regras de Associação Temporais**: tarefa que consiste em encontrar regras de associação entre períodos adjacentes de

tempo. Por exemplo, a regra "$C_1 \rightarrow C_2$" significa que se o conceito $C_1$ aparece no tempo *t*, então o conceito $C_2$ aparece no tempo *t+1*, com determinado nível de suporte e confiança.

**Visualização de Tendências:** consiste em analisar a distribuição de conceitos e seus relacionamentos através de múltiplos subconjuntos de documentos no tempo. Pode ser utilizada para identificar graficamente conceitos estão crescendo ou diminuindo em volume ao longo do tempo. Pode-se também enfatizar os relacionamentos entre conceitos. Nesse caso, permite-se ao usuário ver graficamente as mudanças nos relacionamentos entre conceitos no tempo. É ainda possível comparar grafos/redes de diferentes períodos de tempo (*Trend Graph*).

Nas seções seguintes três das tarefas acima são explicadas em detalhes. Estas três tarefas são: geração de vetores de contexto, descoberta ABC e visualização de tendências.

### 3.7.1 Geração de Vetores de Contexto

Esta tarefa consiste em gerar um vetor de contexto para cada conceito do domínio de análise. Trata-se de um vetor de conceitos ordenados pelo peso da correlação, obtido diretamente do repositório de temas. Além do contexto do conceito, é possível ainda obter o contexto temporal de determinado conceito. Por exemplo, dadas as matrizes de correlação (Figura 15) e correlação temporal (Figura 17), é possível extrair os contextos do conceito "Inovação" em cada um dos quatro anos da fonte de informação, e o contexto agregado (Figura 23).

| **Inovação** ↳ | **2005** | **2006** | **2007** | **2008** | **Agregado** |
|---|---|---|---|---|---|
| | Tecnologia | Sistema | – | Redes | Tecnologia |
| | Conhecimento | Conhecimento | | Tecnologia | Redes |
| | Redes | Informação | | Conhecimento | Conhecimento |
| | Informação | | | Ciência | Informação |
| | Sistema | | | Gestão | Sistema |
| | | | | Informação | Gestão |
| | | | | | Ciência |

Figura 23 – Os cinco conceitos mais relacionados ao conceito "Inovação" classificados em ordem decrescente pelo peso de correlação. Divididos por ano e sem considerar a dimensão tempo (agregado).

Algumas perguntas que podem ser respondidas a partir das informações do modelo: Qual é o vetor de contexto do conceito "Inovação"? Qual era o contexto de "Inovação" *em* 2006? Qual é o contexto de "Inovação" *a partir de* 2006?

### 3.7.2 Descoberta ABC

Como mostrado na revisão da literatura do Capítulo 2, a área de DBL apresenta o modelo de descoberta ABC, que consiste em analisar os relacionamentos indiretos entre conceitos. Uma das formas de se realizar esse tipo de descoberta é através da comparação de seus vetores de contexto (VAN HAAGEN; 'T HOEN et al., 2009). Os dois tipos de descoberta ABC (fechada e aberta), são apresentados a seguir.

a) Descoberta Fechada

Considere a situação na qual se deseja analisar os relacionamentos entre os conceitos "Inovação" ($C_5$) e "Metodologia" ($C_9$) – listados Figura 12 – com base no *corpus* usado como exemplo neste modelo. Ao se verificar a matriz de correlação da Figura 15, na célula i=5 e j=9, obtém-se o valor zero. Isso significa que tais conceitos não coocorrem na fonte de informação analisada. Contudo, é possível utilizar a Descoberta ABC para tentar encontrar relacionamentos indiretos entre os dois conceitos.

Considere A = "Inovação" ($C_5$) e C = "Metodologia" ($C_9$). Verificando a matriz de associação (Figura 20), a célula (5, 9) possui valor igual a 0,001316. Assim, por ser maior que $0^{12}$, os dois conceitos possuem algum relacionamento indireto. Comparando os seus vetores, obtém-se a lista de conceitos em comum que os conectam indiretamente (Figura 24).

12 Pode-se também definir um limiar (*threshold*) com valor maior que zero para uma determinada análise.

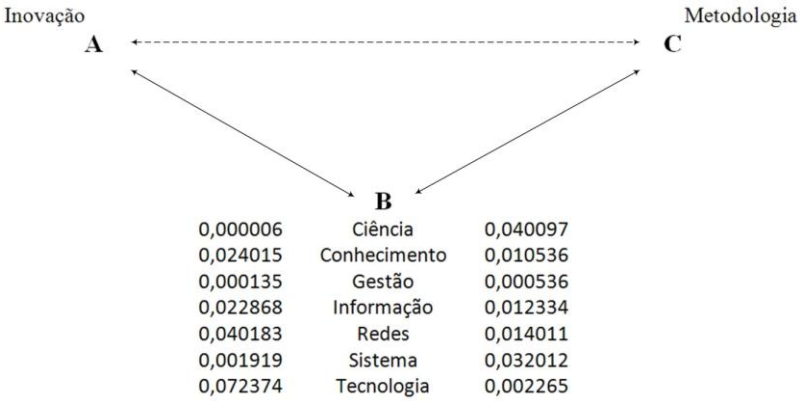


Figura 24 – Descoberta ABC fechada para os conceitos "Inovação" (A) e "Metodologia" (C), e os conceitos que os conectam indiretamente (B).

Calculando a média aritmética simples dos pesos de cada um dos conceitos apresentados na Figura 24, é possível ordenar a lista de conceitos (B) que conectam os conceitos "Inovação" (A) e "Metodologia" (C). Essa lista é apresenta em ordem decrescente de importância na Figura 25. Como pode ser visto, o conceito "Tecnologia" é a principal ligação entre "Inovação" e "Metodologia".

| **Conceitos B** | **(\|A↔B\| + \|B↔C\|)/2** |
|---|---|
| Tecnologia | 0,0373195 |
| Redes | 0,0270970 |
| Ciência | 0,0200515 |
| Informação | 0,0176010 |
| Conhecimento | 0,0172755 |
| Sistema | 0,0169655 |
| Gestão | 0,0003355 |

Figura 25 – Lista em ordem decrescente de importância dos conceitos (B) que conectam "Inovação" (A) e "Metodologia" (C).

b) Descoberta Aberta

A descoberta aberta parte apenas de um conceito (A). Assim, considere a situação na qual se deseja buscar conceitos que se ligam indiretamente ao conceito "Inovação" ($C_5$). O primeiro passo é buscar o

vetor de “Inovação” a partir da matriz de correlação (Figura 15). Estando esse vetor ordenado em ordem decrescente pelo valor da correlação, devem-se escolher os *k* primeiros elementos (excluindo o próprio conceito) para serem considerados os conceitos intermediários (B). Para cada conceito em B, deve-se buscar o vetor na matriz de correlação. É necessário excluir desse vetor o conceito A e os conceitos que estão em B. O vetor resultante, ordenado, terá os conceitos que se conectam indiretamente (C) ao conceito “Inovação”.

| | |
|---|---|
| Inovação | 1 |
| Tecnologia | 0,072374 |
| Redes | 0,040183 |
| Conhecimento | 0,024015 |
| Informação | 0,022868 |
| Sistema | 0,001919 |
| Gestão | 0,000135 |
| Ciência | 0,000006 |

Figura 26 – Vetor de Contexto do conceito “Inovação”.

Usando *k*=1, tem-se o seguinte conceito intermediário (B): “Tecnologia”. O próximo passo é buscar o vetor de “Tecnologia” (Figura 27).

| | |
|---|---|
| Tecnologia | 1 |
| Informação | 0,101448 |
| Inovação | 0,072374 |
| Gestão | 0,036050 |
| Conhecimento | 0,035360 |
| Redes | 0,028597 |
| Ciência | 0,017865 |
| Sistema | 0,008498 |
| Metodologia | 0,002265 |
| Qualidade | 0,001426 |

Figura 27 – Vetor de Contexto do conceito “Tecnologia”.

Deste vetor, tira-se o próprio conceito (tecnologia) e os conceitos do vetor de “Inovação” (A), apresentado na Figura 26 (estes estão ligados à “Inovação” diretamente). Os conceitos restantes foram “Metodologia” e “Qualidade”, que formam o elemento C, da tríplice

ABC. Assim, pode-se dizer que o conceito “Inovação” (A) está ligado indiretamente aos conceitos “Metodologia” e “Qualidade” (C), por intermédio de “Tecnologia” (B). Caso *k* fosse maior que 1, o mesmo processo teria de ser feito para os demais conceitos em B.

### 3.7.3 Visualização de Tendências

A partir das informações sobre as frequências temporais dos conceitos é possível ver graficamente a distribuição dessas frequências na fonte de informação ao longo do tempo. A Figura 28 mostra essa situação para os conceitos “Redes”, “Gestão” e “Inovação”. A partir do gráfico é possível ver uma queda acentuada nas frequências dos conceitos “Redes” e “Inovação” no ano de 2007 e uma leve alta em 2008. Já o conceito “Gestão” mostra uma alta consistente na sua frequência ao longo de 2005 a 2008.

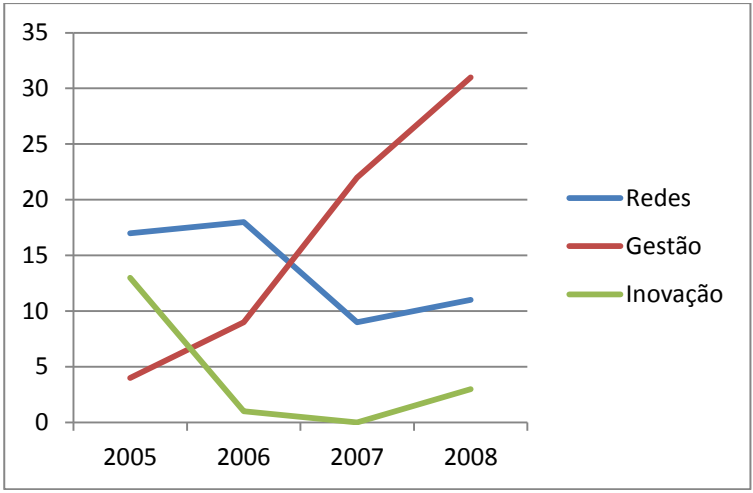


Figura 28 – Distribuição da frequência dos conceitos “Redes”, “Gestão” e “Inovação” ao longo do tempo.

E a partir das informações presentes no repositório de temas de análise, é possível ver graficamente a distribuição do peso de uma relação entre dois conceitos. Por exemplo, a Figura 29 apresenta a

distribuição do peso de correlação e associação entre os conceitos “Ciência” e “Redes”.

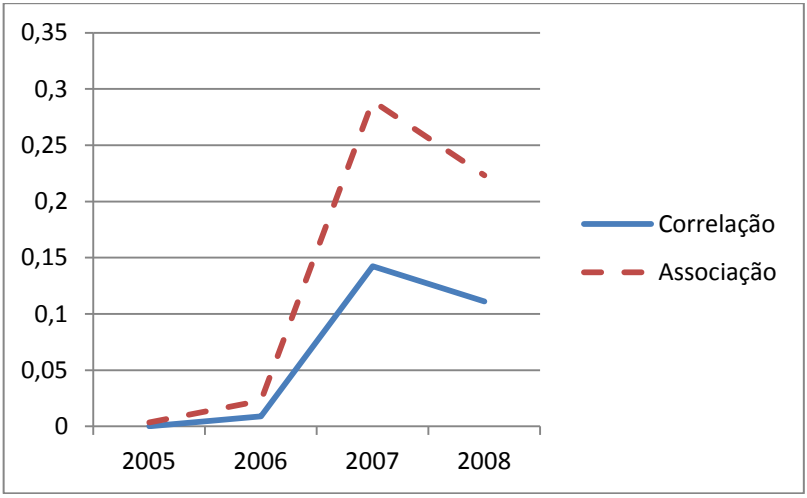


Figura 29 – Distribuição do peso da relação entre os conceitos “Ciência” e “Redes”.

É possível ver que em 2005 a correlação entre os dois conceitos é zero. Isso quer dizer que os dois conceitos ainda não coocorrem nesse ano. Contudo, é possível ver que há um relacionamento indireto entre eles, pois, a associação entre eles nesse ano é 0,003197. Para saber quais conceitos conectam “Ciência” e “Redes” em 2005 é necessário utilizar a Descoberta ABC, apresentada em detalhes na próxima seção.

## 3.8 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o modelo de *Temporal Knowledge Discovery in Texts* (TKDT) proposto neste trabalho. Este modelo é dividido por fases, assim como os modelos tradicionais de descoberta de conhecimento. As fases deste modelo são: configuração dos temas de análise, identificação das ocorrências dos conceitos, correlação e correlação temporal, associação e associação temporal, criação do repositório de temas de análise, e tarefas intensivas em conhecimento, com ênfase nos relacionamentos diretos e indiretos entre os conceitos do

domínio. Cada uma destas fases foi explicada em detalhes utilizando-se como exemplo uma fonte de informações não estruturadas e com um atributo temporal e um conjunto de instâncias de uma ontologia como domínio de análise. Na fase de tarefas intensivas em conhecimento, as tarefas de *geração de vetores de contexto* e *descoberta abc* foram examinadas em detalhes. Vale lembrar que a lista de tarefas de conhecimento apresentada não é exaustiva. O próximo capítulo apresenta uma implementação de um protótipo de acordo com o modelo proposto.

# 4 PROTÓTIPO BASEADO NO MODELO PROPOSTO

Este capítulo apresenta um protótipo de um sistema baseado no modelo de TKDT proposto neste trabalho, que é descrito no capítulo anterior. É apresentada a arquitetura desse protótipo e como cada um de seus módulos implementa parte do modelo proposto.

## 4.1ARQUITETURA DO PROTÓTIPO

A Figura 30 apresenta a arquitetura do protótipo baseado no modelo de TKDT proposto neste trabalho. Os módulos dentro do quadro pontilhado representam o núcleo do protótipo, que funciona como um *framework* sobre o qual serviços de conhecimento são construídos.

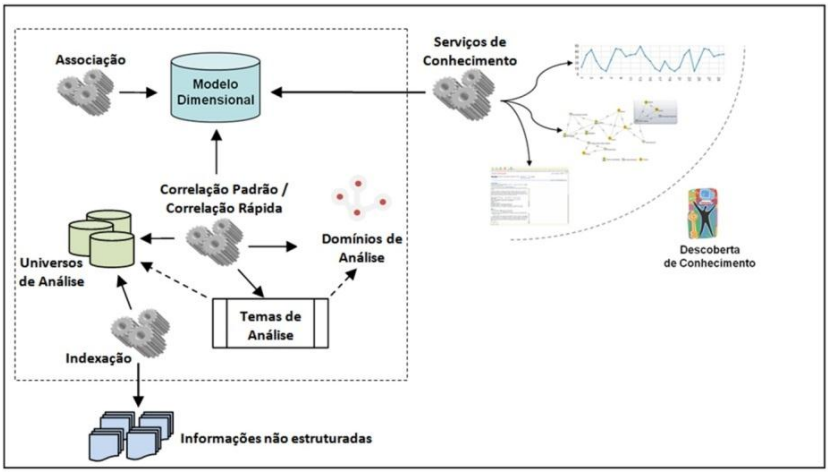


Figura 30 – Arquitetura do protótipo do modelo de TKDT.

A seguir descrevem-se os principais componentes do sistema.

**Configuração dos Temas de Análise**: Para cada tema de análise, o usuário deve informar os conceitos do domínio de análise e as fontes de informação para o universo de análise. Para os conceitos, é necessário fornecer os seus nomes e respectiva classe (por exemplo, conceito: *Ciência*, classe: *Keyword*). Para o universo de análise é

necessário informar os índices invertidos das respectivas fontes de informação.

**Indexação das Fontes de Informação**: este módulo é responsável pela geração de um índice invertido para cada fonte de informação utilizada nas análises. A indexação da fonte de informação é necessária para os passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. Este módulo é explicado em mais de detalhes na seção 4.2.

**Correlação Padrão**: este módulo realiza os passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. Para cada documento indexado realiza-se o seguintes passos: contam-se as ocorrências e coocorrências dos conceitos do domínio de análise; obtêm-se o *timestamp* do documento, quando estiver disponível; e armazena-se essas informações no Repositório de Temas de Análise. Depois que todos os documentos foram analisados, é realizado o cálculo da correlação e correlação temporal.

**Correlação Rápida**: este módulo também realiza os passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. Basicamente, são realizadas consultas ao índice invertido a partir de conceitos e pares de conceitos do domínio de análise com objetivo de identificar quais conceitos estão na fonte de informação, suas frequências individuais e conjuntas e os seus *timestamps*. Este módulo é explicado em mais detalhes na seção 4.3.

**Associação**: este módulo é responsável pelo cálculo das matrizes de associação e associação temporal. O cálculo de similaridade entre os vetores foi realizado utilizando-se a função cosseno, como foi apresentado na seção 2.4.2.

**Modelo dimensional**: além dos módulos citados acima, o protótipo utiliza-se de um modelo dimensional para se armazenar o Repositório de Temas de Análise. Tal abordagem também se deve a requisitos de desempenho. Na seção 4.4 descreve-se em detalhes o modelo de dados utilizado.

**Serviços de conhecimento**: cada serviço de conhecimento implementado no protótipo é composto por uma ou mais tarefas intensivas em conhecimento apresentadas no modelo. Esse módulo possui até o momento os serviços *Perfil de Conceitos* e *Redes de Relacionamentos*. O serviço *Perfil de Conceitos* é uma implementação direta da tarefa *Geração de Vetores de Contexto*. O perfil de um determinado conceito é representado por um vetor com os conceitos

mais fortemente conectados a ele baseado na força de correlação. Pode-se obter o perfil de forma agregada (desconsiderando-se o tipo de relação e o tempo) ou combinando-se os tipos de relações e o tempo. O serviço *Redes de Relacionamentos* possibilita a visualização de uma rede de conceitos a partir de um determinado conceito de interesse informado pelo usuário (tarefa *Visualização de Tendências*). Através de uma rede de relacionamentos torna-se possível ter uma visão mais ampla de como conceitos de um determinado domínio de aplicação se conectam entre si. Os relacionamentos podem ser definidos por relações diretas entres os conceitos (tarefa *Geração de Vetores de Contexto*) ou relações indiretas (tarefa *Descoberta ABC*).

## 4.2 INDEXAÇÃO DAS FONTES DE INFORMAÇÃO

O primeiro passo para uso de uma fonte de informação é a sua indexação utilizando-se métodos da área de RI. Assim, o objetivo deste módulo é a geração de um índice para cada fonte de informação. No modelo proposto, além das informações normalmente presentes em um índice invertido, é necessário que armazenar a informação temporal dos documentos. Logo, para a indexação das fontes de informação, optou-se por utilizar a estrutura de índice invertido apresentada conceitualmente na Figura 31.

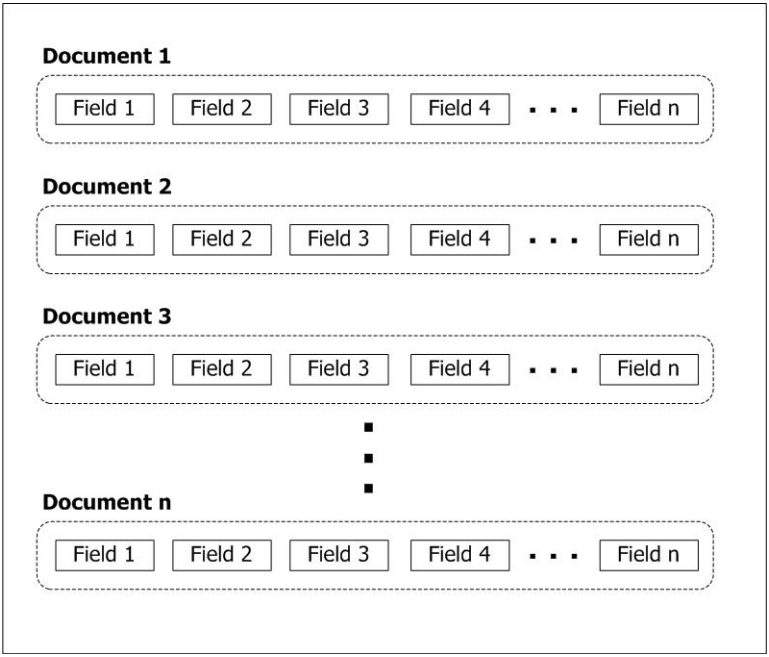


Figura 31 – Representação conceitual de um índice textual.

Como pode ser visto, um índice é representado por uma sequência de *Documents*, no qual cada *Document* possui um conjunto de *Fields*. Cada *Field* tem um nome e um valor textual. Um *Field* pode ser *indexado* ou apenas *armazenado*. Os *Fields* indexados são utilizados para se construir a lista de termos do índice invertido (como mostrado no Capítulo 2) que são utilizados para a busca. Já no caso de *Fields* armazenados, o texto inteiro é armazenado para posterior recuperação. Assim, o módulo de indexação do protótipo possui um índice invertido, no qual cada *Document* possui três *Fields*: um *Field* indexado com o conteúdo textual do documento, um *Field* armazenado com o identificador do documento, e outro *Field* armazenado com o *timestamp* do documento. Deste modo, é possível realizar uma busca textual sobre o índice para recuperar os documentos que contêm o termo buscado e seus respectivos *timestamps.* A Figura 32 apresenta um exemplo de índice para três documentos.

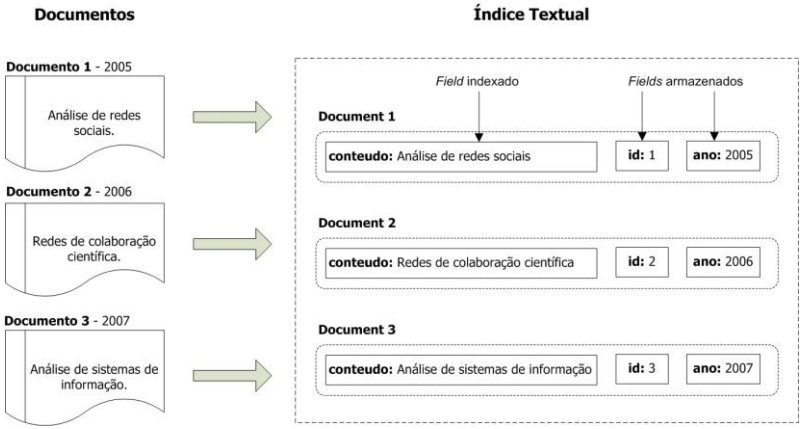


Figura 32 – Exemplo de um índice textual para três documentos.

O índice com informação temporal gerado nessa fase é utilizado para se descobrir as frequências individuais e conjuntas e o *timestamp* dos conceitos do domínio de análise estão presentes no universo de análise.

## 4.3 CORRELAÇÃO RÁPIDA

Este módulo é responsável pelos passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. Nesta abordagem, as frequências individuais e conjuntas e os *timestamps* dos conceitos do domínio de análise são obtidas através de consultas (*queries*) ao índice da fonte de informação. Para cada par de conceitos $C_1$ e $C_2$ do domínio de análise têm-se três consultas: "$C_1$" (frequência individual do conceito $C_1$), "$C_2$" (frequência individual do conceito $C_2$) e "$C_1$ AND $C_2$" (frequência conjunta dos conceitos $C_1$ *e* $C_2$). Para cada uma das três consultas ao índice obtêm-se como resposta uma lista de documentos e o seus respectivos *timestamps*. O número de documentos retornado é utilizado como frequência do conceito (ou do par de conceitos) e o *timestamp* é utilizado para se calcular as frequências (individual e conjunta) temporais.

Assim, para cada documento no qual um conceito (ou par de conceitos) aparece, considera-se apenas uma ocorrência (ou coocorrência) do conceito naquele documento. A Tabela 11 ilustra essa

situação para os conceitos “Ciência” e “Informação” (coluna 1). Neste caso três consultas são realizadas: “Ciência”, “Informação” e “Ciência AND Informação” (coluna 2). O número de documentos recuperados representam as frequências totais e por ano (coluna 3).

| **Par de Conceitos** | **Tipo de frequência** | **Consultas** | **Número documentos recuperados** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **Total** | **Dividido por ano** | | | |
| | | | | **2005** | **2006** | **2007** | **2008** |
| $C_1$: Ciência<br>$C_2$: Informação | Individual | “Ciência” | 23 | 5 | 7 | 9 | 2 |
| | | “Informação” | 45 | 6 | 11 | 22 | 5 |
| | Conjunta | “Ciência AND Informação” | 15 | 1 | 5 | 6 | 3 |

Tabela 11 – Consultas ao índice textual utilizando-se um par de conceitos do domínio de análise. O número de documentos recuperados é utilizado como frequência (individual e conjunta).

Com as informações apresentadas na Tabela 11 tem-se as frequências individuais e conjuntas (totais e por ano) e, com o número de documentos presentes no índice (tamanho do *corpus*), calcula-se a correlação e a correlação temporal utilizando-se o *Phi-Squared*[13], como foi apresentado no capítulo anterior.

Assim, os conceitos “Ciência” e “Informação”, suas frequências individuais e conjuntas (totais e divididas por ano), e os pesos de correlação e correlação temporal são armazenados no Repositório de Temas de Análise, que neste protótipo é representado por um modelo dimensional de dados, apresentado na próxima seção.

## 4.4 MODELO DIMENSIONAL

Para representação de dados no protótipo utilizou-se conceitos de *Data Warehouse* (DW). Segundo Inmon (1997), DW é “um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”. Dessa definição, têm-se dois aspectos relacionados ao modelo proposto: “dados baseados em

13 Este método foi escolhido por fornecer valores já normalizados entre 0 e 1 e por apresentar bom desempenho em tarefas de correlação (GONÇALVES, 2006).

assuntos", que corresponde aos temas de análise do modelo; e "variável em relação ao tempo", que corresponde à dimensão tempo do repositório do modelo. Além disso, DWs são modelados para permitir que consultas diversas sejam processadas com alto desempenho pelas ferramentas analíticas. O modelo de dados normalmente utilizado na construção de DWs é o modelo dimensional (KIMBALL; REEVES et al., 1998; GIOVINAZZO, 2000; KIMBALL; ROSS, 2002).

Portanto, os conceitos (domínio de análise) que foram extraídos das fontes de informação (universo de análise) e o Repositório de Temas de Análise são representadas no protótipo pelo modelo dimensional ilustrado na Figura 33.

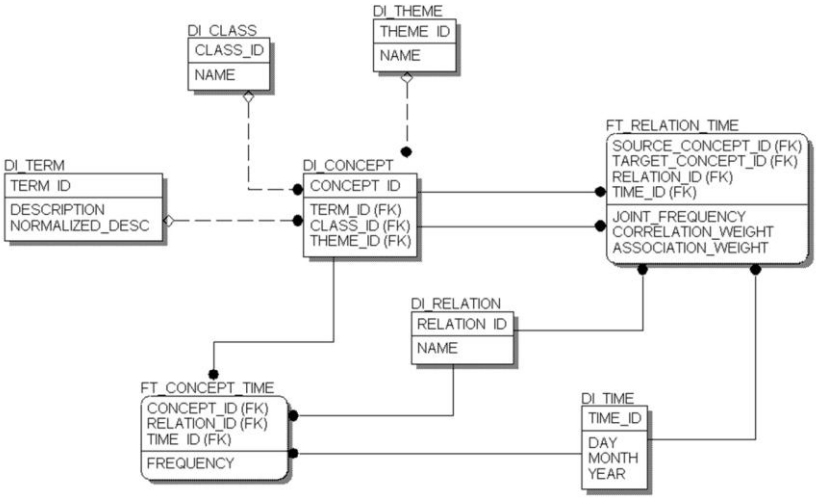


Figura 33 – Modelo dimensional utilizado no protótipo do modelo.

As dimensões DI_TERM, DI_CLASS e DI_CONCEPT são utilizadas para representar os conceitos dos domínios de análise:

**DI_TERM**: dimensão utilizada para representar as instâncias de classes de diferentes domínios de análise. Essa dimensão apenas representa a descrição textual de um termo. Ela pode pertencer a diferentes classes em diferentes temas de análise. Possui atributos para a

descrição do termo (*description*) e descrição normalizada[14] (*normalized_desc*).

**DI_CLASS**: utilizada para representar as classes dos conceitos do domínio de análise. Cada classe é identificada por um número sequencial (*class_id*). Também possui um atributo que descreve o nome da classe (*name*).

**DI_CONCEPT**: representa os diversos conceitos de um tema de análise. Esses conceitos são instâncias do domínio de análise, que possuem uma descrição textual (*term_id*), uma classe (*class_id*), e pertencem a um determinado tema (*theme_id*). Cada conceito é identificado por um número sequencial (*concept_id*).

Já as dimensões DI_THEME, DI_RELATION, DI_TIME e tabelas de fato FT_CONCEPT_TIME e FT_RELATION_TIME são utilizadas para armazenar as informações do repositório de temas de análise, representado conceitualmente pela ontologia apresentada na Figura 22:

**DI_THEME**: corresponde à classe *Theme*. Dimensão utilizada para armazenar os temas de análise. Cada tema é representado por um número sequencial (*theme_id*) e por um nome (*name*).

**DI_RELATION**: corresponde à classe *Relation*. Representa os tipos de relações entre os conceitos. Cada relação tem um identificador (*relation_id*) e um nome (*name*).

**DI_TIME**: corresponde à classe *Time*. Utilizada para representar o tempo associado aos conceitos e seus relacionamentos. Possui um sequencial (*time_id*) como identificador, e representa a dimensão tempo em cinco granularidades diferentes: dia (*day*), mês (*month*) e ano (*year*).

**FT_CONCEPT_TIME**: corresponde às classes *ConceptEvent* e *TemporalConceptEvent*. Representa a ocorrência individual de um determinado conceito (*concept_id*) em uma determinada relação (*relation_id*), em determinado tempo (*time_id*), e em um determinado tema (*theme_id*). A frequência é representada pela propriedade *frequency*.

**FT_RELATION_TIME**: corresponde às classes *RelationEvent* e *TemporalRelationEvent*. Representa a ocorrência conjunta de dois conceitos (*source_concept_id* e *target_concept_id*) em um determinado

14 A normalização refere-se ao processo de reduzir um termo à sua raiz. Por exemplo, os termos “tecnologia” e “tecnologias”, serão reduzidos para apenas um termo: “tecnolog”.

tipo de relação (*relation_id*), em determinado tempo (*time_id*), e em um determinado tema (*theme_id*). Além da frequência conjunta (*joint_frequency*), contêm a correlação (*correlation_weight*) e a associação (*association_weight*) entre os dois conceitos.

## 4.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou um protótipo de um sistema desenvolvido de acordo com o modelo de *Temporal Knowledge Discovery in Texts* (TKDT) proposto neste trabalho. Esse protótipo possui um módulo que permite aos usuários a configuração de temas de análise; um módulo de indexação das fontes de informação; dois módulos de correlação, chamados de "Correlação Rápida" e “Correlação Padrão”, que são responsáveis pelas fases de identificação das ocorrências dos conceitos, correlação e correlação temporal; um módulo para cálculo da força de associação entre os conceitos; e um módulo de serviços de conhecimento. As informações do repositório de temas de análise são mapeadas em um modelo de dados dimensional. O serviços de conhecimento implementados são: *Perfil de Conceitos* e *Redes de Relacionamentos*.

# 5 DEMONSTRAÇÃO DE VIABILIDADE E ANÁLISE COMPARATIVA

Este capítulo está dividido em duas partes. A primeira descreve um estudo de caso para demonstração de viabilidade do modelo proposto. O cenário de aplicação é apresentado bem como os serviços de conhecimento utilizados: *Perfil de Conceitos* e *Redes de Relacionamentos*. A segunda parte do capítulo descreve uma análise comparativa do modelo proposto com o modelo proposto por Gonçalves (2006). O objetivo é discutir as contribuições da tese à área de descoberta de conhecimento em textos.

## 5.1 CENÁRIO DE APLICAÇÃO

Este estudo de caso utiliza como *universo de análise* uma base com informações de currículos da Plataforma Lattes[15], em formato XML, de aproximadamente 1.000 pesquisadores da Universidade Federal de Santa Catarina (UFSC). O Currículo Lattes de um pesquisador possui, entre outros, itens relativos a sua produção científica, formação acadêmica e atividade profissional. Cada um desses itens possui um conjunto de palavras-chave informado pela própria pessoa.

O *domínio de análise* é composto por dois tipos de conceitos: (a) classe *Pessoa*, que possui um identificador e nome das pessoas das quais são utilizadas seus currículos, e (b) classe *PalavraChave*, que representa as palavras-chave referentes a produção científica, formação acadêmica e atividade profissional de cada pessoa.

15 http://lattes.cnpq.br/

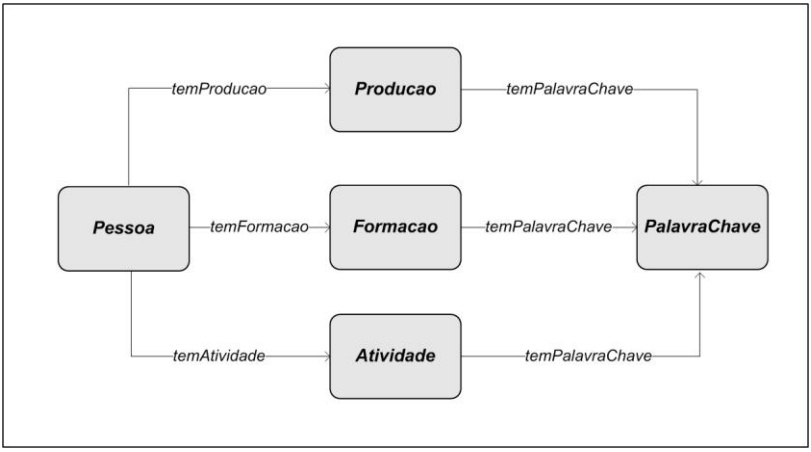


Figura 34 – Ontologia que descreve o domínio de análise do estudo de caso.

O objetivo é utilizar os relacionamentos diretos (correlação) entre conceitos *Pessoa* e conceitos *PalavraChave*, distribuídos ao longo do tempo, para construção de perfis (serviço *Perfil de Conceitos*). Além disso, os relacionamentos indiretos (associação) entre conceitos do tipo *Pessoa*, ao longo do tempo, serão utilizados para visualização de redes (serviço de *Redes de Relacionamentos*).

Além da dimensão temporal dos perfis e das redes, pretende-se criar diferentes perfis e diferentes redes para cada pessoa de acordo com diferentes pesos dados a cada parte do currículo. Por exemplo, para o contexto *Docente*, itens do currículo relativos a ensino podem ter um peso maior do que artigos publicados. Já para o contexto *Pesquisador*, pode ocorrer o oposto. Os diferentes tipos de coocorrência (ensino, artigo, etc.) entre um conceito do tipo *Pessoa* e os conceitos do tipo *PalavraChave* (informados em seu currículo) serão utilizados para calcular os relacionamentos em cada contexto. Os tipos de contextos utilizados para cada currículo do universo de análise são:

- *Pesquisador*: este contexto atribui mais peso para itens do currículo dos seguintes tipos: atividades de pesquisa e desenvolvimento, publicação de artigos em periódicos e anais de eventos, relatórios de pesquisa, orientações;
- *Gestor*: privilegia itens relacionados a atividades de direção e administração, e participação em conselhos, comissões e consultoria;

- *Extensionista*: contexto que atribui mais peso para itens relacionados extensão universitária, trabalhos técnicos e relatório de pesquisa;
- *CompetenciaInovacao*: para este contexto os itens mais importantes são: pesquisa e desenvolvimento, participação em conselhos, comissões e consultoria, artigo publicado em periódicos, software, livro ou capítulo de livro;
- *Docente*: neste contexto, os itens de maior peso são as atividades relacionadas ao ensino;
- *Geral*: para este tipo de contexto todos os itens do currículo possuem o mesmo peso.

A Figura 35 apresenta um exemplo de como são gerados os relacionamentos para dois contextos (*Docente* e *Pesquisador*) para uma determinada pessoa. Tem-se o currículo de uma pessoa (chamado apenas de João, para simplificar), com itens relativos à sua atividade profissional (ensino, pesquisa e desenvolvimento) e à sua produção científica (artigos e livros). São calculados as frequências e os pesos de correlação entre o conceito *João* e os conceitos p1, p2, p3, p4 e p5 (*PalavraChave*), de acordo com os quatro diferentes tipos de coocorrência (ensino, pesquisa e desenvolvimento, artigos, e livros), e de acordo com a informação temporal disponível (ano). Os contextos *Pesquisador* e *Ensino* são calculados a partir dos pesos dados aos tipos de coocorrência.

O processo descrito na Figura 35 é realizado para todos os currículos, utilizando todas as palavras-chave presentes em itens de produção científica, formação e atuação profissional, para todos os seis contextos descritos anteriormente. O peso de cada item em cada contexto está definido em um arquivo XML.

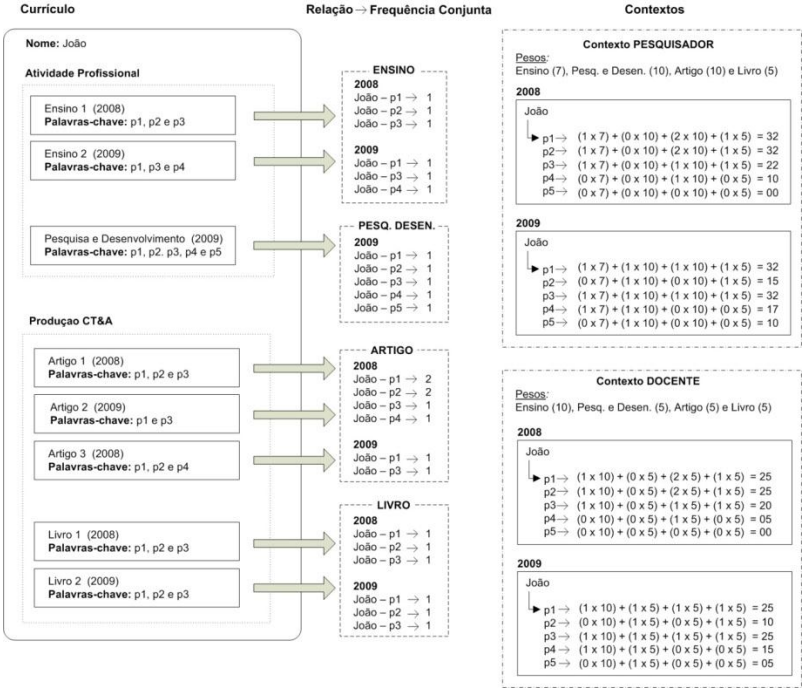


Figura 35 – Exemplo simplificado de um currículo, os tipos de coocorrência e o cálculo dos relacionamentos para os contextos *Pesquisador* e *Docente*.

Por fim, o repositório de temas de análise (apresentado na Figura 22) utilizado nesse estudo de caso contém os seguintes itens:

- Temas de análise (*Theme*): *Pesquisador*, *Gestor*, *Extensionista*, *CompetenciaInovacao*, *Docente* e *Geral*. Cada contexto foi modelado como um tema;
- Relações (*Relation*): *Perfil*, para representar os relacionamentos entre pessoas e palavras-chave, baseado nos relacionamentos diretos (correlação), e *Rede*, para representar os relacionamentos entre pessoas, baseado nos relacionamentos indiretos (associação);
- Tempo (*Time*): ano de produção, formação ou atuação profissional;
- Ocorrência de conceitos (*ConceptEvent*): número de currículos utilizados (classe *Pessoa*), e número do palavras-chave (classe

*PalavraChave*) distintas encontradas nos itens de produção, formação ou atuação profissional;

- Ocorrência de conceitos no tempo (*TemporalConceptEvent*): número do palavras-chave (classe *PalavraChave*) distintas encontradas nos itens de produção, formação ou atuação dos currículos em cada ano (*Time*);
- Ocorrência de relações (*RelationEvent*): entre *Pessoa* e *PalavraChave* (relação do tipo *Perfil*) e entre *Pessoa* e *Pessoa* (relação do tipo *Rede*);
- Ocorrência de relações no tempo (*TemporalRelationEvent*): entre *Pessoa* e *PalavraChave* (relação do tipo *Perfil*) e entre *Pessoa* e *Pessoa* (relação do tipo *Rede*) em cada ano (*Time*);

## 5.2 SERVIÇO *PERFIL DE CONCEITOS*

O perfil de um determinado conceito é representado por um vetor com os conceitos mais fortemente conectados a ele baseado na força de correlação. Pode-se obter o perfil temporal ou o perfil de forma agregada (desconsiderando-se o tempo). Como no estudo de caso em questão os relacionamentos entre os conceitos da classe *Pessoa* e da classe *PalavraChave* mudam de acordo com o contexto (tema), cada conceito possui seis diferentes perfis (*Pesquisador*, *Gestor*, *Extensionista*, *CompetenciaInovacao*, *Docente* e *Geral*).

Para exemplificar, apresenta-se os diferentes tipos de perfis que podem ser obtidos a partir do currículo de Roberto Carlos dos Santos Pacheco. A Figura 36 apresenta visualmente o perfil do tema/contexto *Geral* de “Pacheco”, sem considerar a dimensão tempo. São apresentados somente os 10 conceitos com maior peso, contudo a lista completa inclui 446 conceitos.

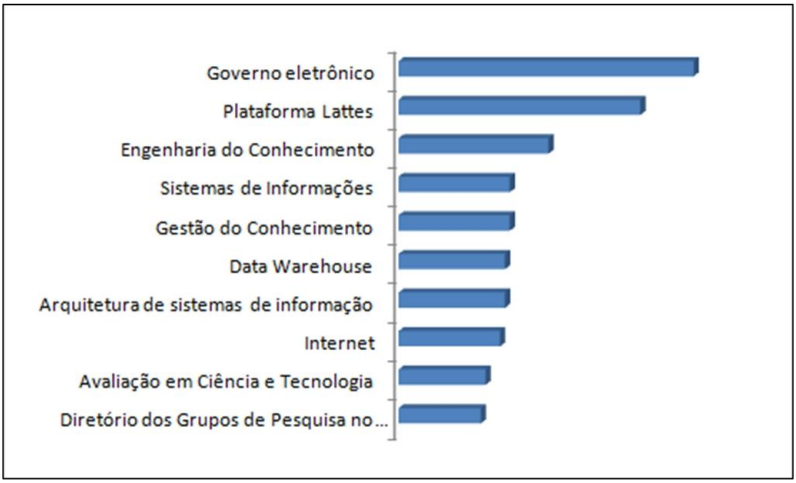


Figura 36 – Perfil do tema *Geral* de "Pacheco" (sem considerar a dimensão tempo).

Já a Figura 37 apresenta o perfil do tema/contexto *Pesquisador* de "Pacheco", também sem considerar a dimensão tempo. Percebe-se que surgiram alguns conceitos novos (ex.: "Inteligência Artificial"), enquanto outros desapareceram (ex.: "Avaliação em Ciência e Tecnologia"). E a relevância dos conceitos que permaneceram no perfil foi alterada, devido à ponderação dada a cada item.

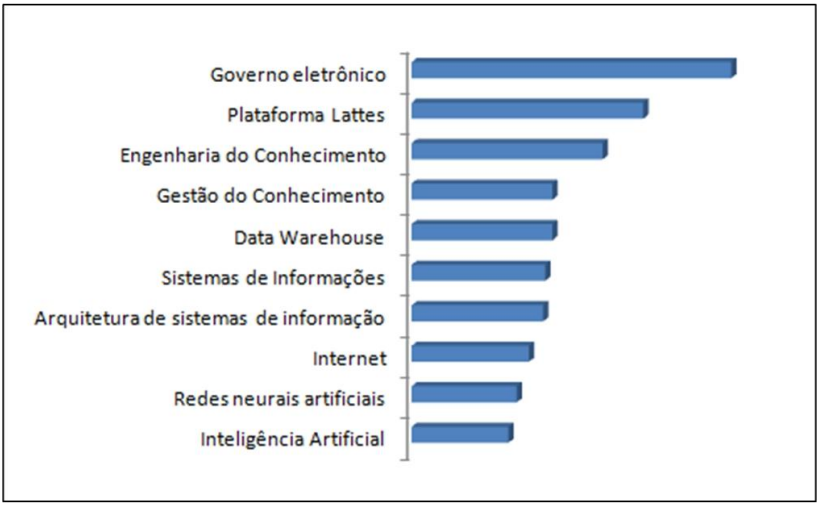


Figura 37 – Perfil do tema *Pesquisador* de "Pacheco" (sem considerar a dimensão tempo).

Os perfis nos outros 4 contextos/temas (*Docente, CompetenciaInovacao, Extensionista e Gestor*) de "Pacheco" são apresentados na Figura 38.

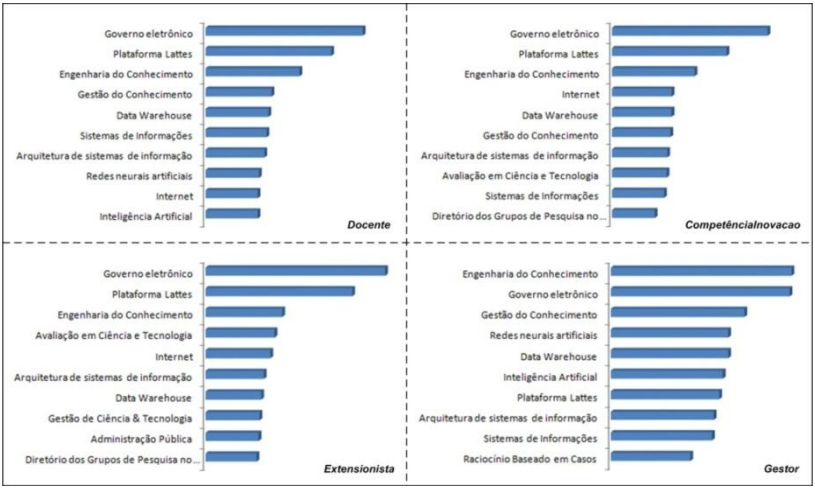


Figura 38 – Perfis dos temas *Docente, CompetenciaInovacao, Extensionista e Gestor* de "Pacheco" (sem considerar a dimensão tempo).

Os perfis de conceitos apresentados até agora não são temporais. Como se tem no currículo o ano de cada palavra-chave, e essa informação foi adicionada ao repositório para o tema de análise em questão, pode-se ver os perfis temporais de cada pessoa. No caso do perfil de "Pacheco", os conceitos estão distribuídos no período que vai de 1984 a 2010.

A partir dos dados do repositório a seguinte pergunta, por exemplo, pode ser respondida: Qual era o perfil do tema *Geral* de "Pacheco" *em* 2006? A Figura 39 mostra a resposta para essa pergunta.

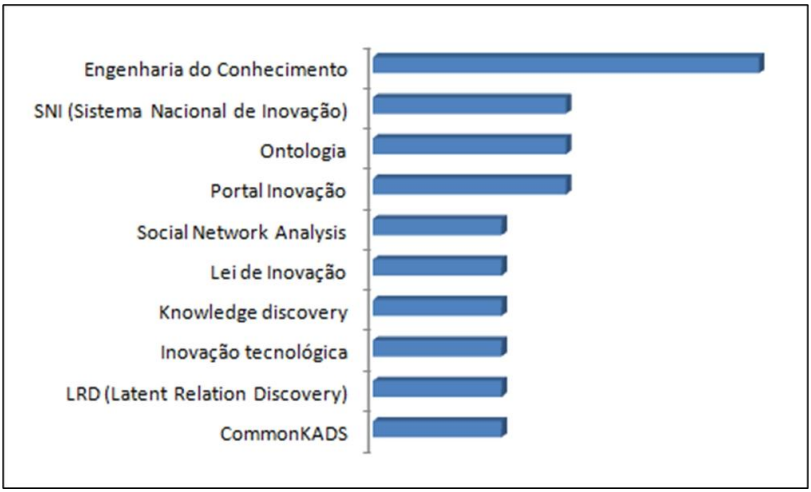


Figura 39 – Perfil do tema *Geral* de "Pacheco" no ano de 2006.

Outro exemplo de pergunta que pode ser respondida: Qual era o perfil *Geral* de "Pacheco" *antes de* 2006 e *a partir de* 2006? Veja a Figura 40.

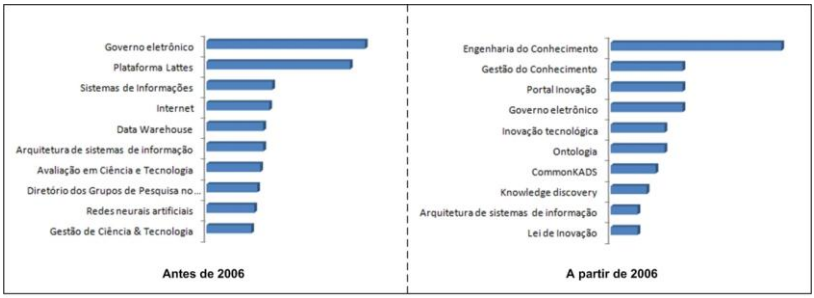


Figura 40 – Perfil do tipo *Geral* de "Pacheco" *antes* e *a partir de* 2006.

É possível também comparar períodos determinados de tempo. Por exemplo, comparar o perfil de "Pacheco" de 1997 a 2003 com o período de 2004 a 2010. Veja a Figura 41.


Figura 41 – Perfil do tema *Geral* de "Pacheco" de 1997 a 2003 e de 2004 a 2010.

Outra forma de ver os relacionamentos temporais entre pessoas e palavras-chave é apresentar a evolução no tempo de determinados conceitos no perfil de uma pessoa. Por exemplo, a Figura 42 mostra a evolução dos conceitos "Governo Eletrônico" e "Engenharia do Conhecimento" no perfil (tema/contexto *Geral*) de "Pacheco". Percebe-se que até o ano de 2004 o conceito "Governo Eletrônico" possuía maior relevância. A partir desse ano esse conceito tem uma queda acentuada e o conceito "Engenharia do Conhecimento" passa a crescer e o ultrapassa em frequência no currículo em análise.

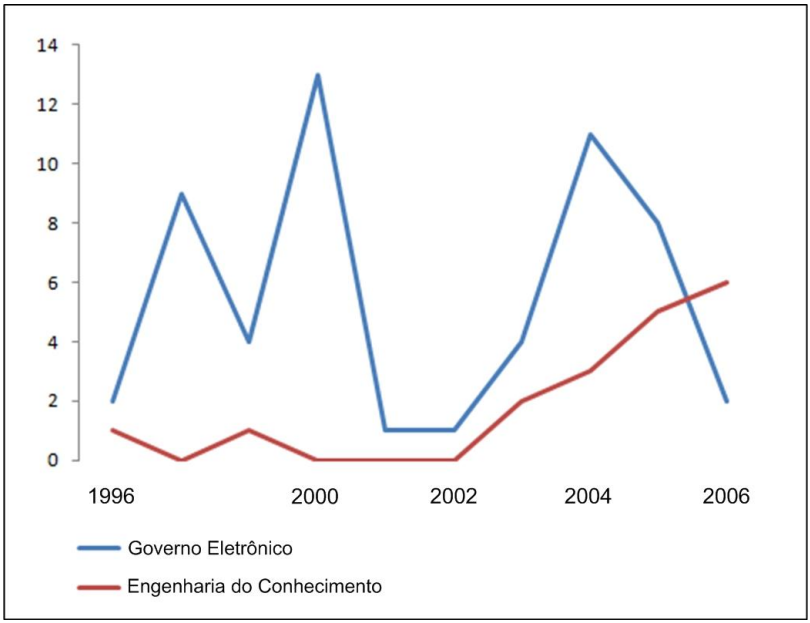


Figura 42 – Conceitos “Governo Eletrônico” e “Engenharia do Conhecimento” no tempo (perfil do tema *Geral* de “Pacheco”).

É importante lembrar que esses perfis são construídos a partir das relações entre os conceitos do tipo *Pessoa* e *PalavraChave* armazenadas no repositório. Os perfis mostrados até agora partem de pessoas. Mas o inverso também é possível: ver os perfis a partir das palavras-chave. Da mesma forma que é feito o *Perfil de Conceitos* de uma pessoa, também é possível ver o *Perfil de Conceitos* de uma palavra-chave. Nesse caso, dada uma palavra-chave, é possível ver as pessoas mais fortemente relacionadas a ela. A Figura 43 mostra o perfil de “Gestão do Conhecimento”.

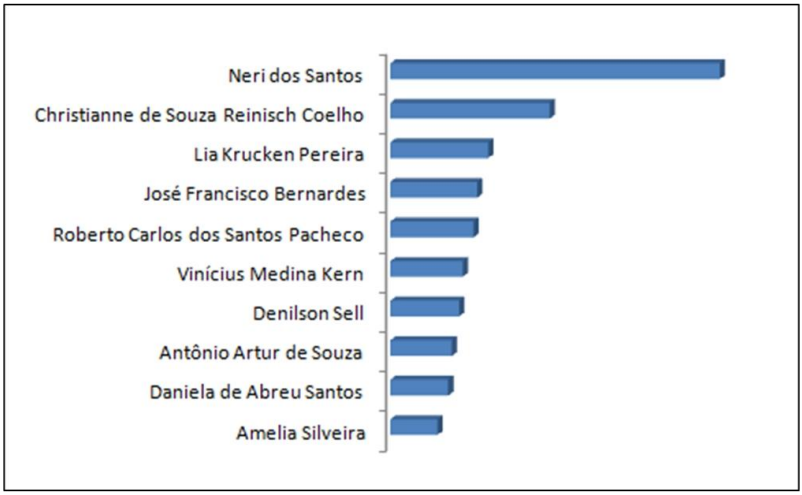


Figura 43 – Perfil (tema *Geral*) do conceito "Gestão do Conhecimento".

Logo, também é possível ver evolução no tempo de pessoas relacionadas a um termo. Por exemplo, a Figura 44 mostra a evolução dos conceitos "Pacheco" e "Kern" no perfil temporal de "Gestão do Conhecimento". Isso também pode ser interpretado como a evolução no tempo do conceito "Gestão do Conhecimento" nos perfis de "Pacheco" e "Kern".

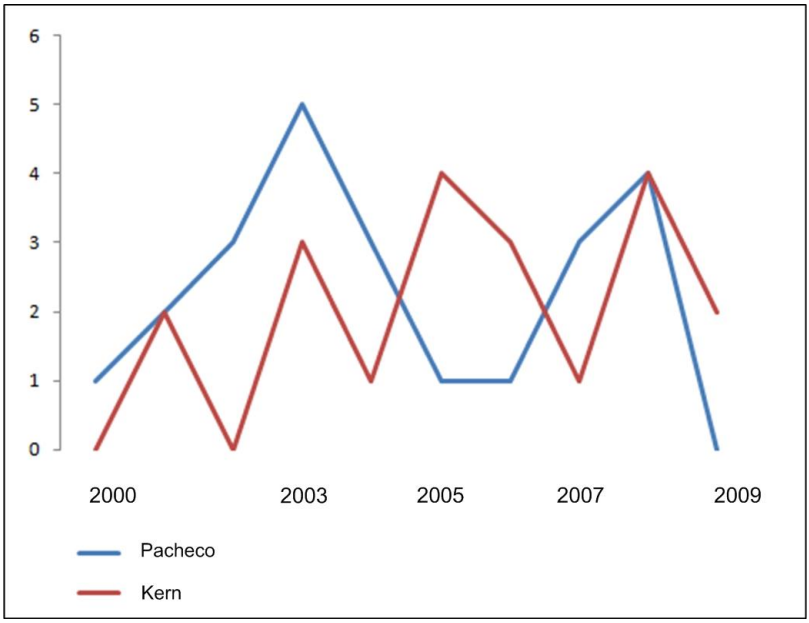


Figura 44 – Conceitos “Pacheco” e “Kern” no tempo (perfil do tema *Geral* de “Gestão do Conhecimento”).

As diferentes formas de se analisar os perfis mostrados até agora são baseadas na correlação entre conceitos da classe *Pessoa* e conceitos da classe *PalavraChave*. Não há no repositório a correlação entre os conceitos da classe *Pessoa*. Contudo, pode-se utilizar a associação. Por exemplo, pode-se projetar graficamente o peso de associação entre dois conceitos ao longo do tempo. Nesse caso, a associação mostra a aderência dos perfis entre as pessoas. A Figura 45 mostra um exemplo para os conceitos “Pacheco” e “Kern”.

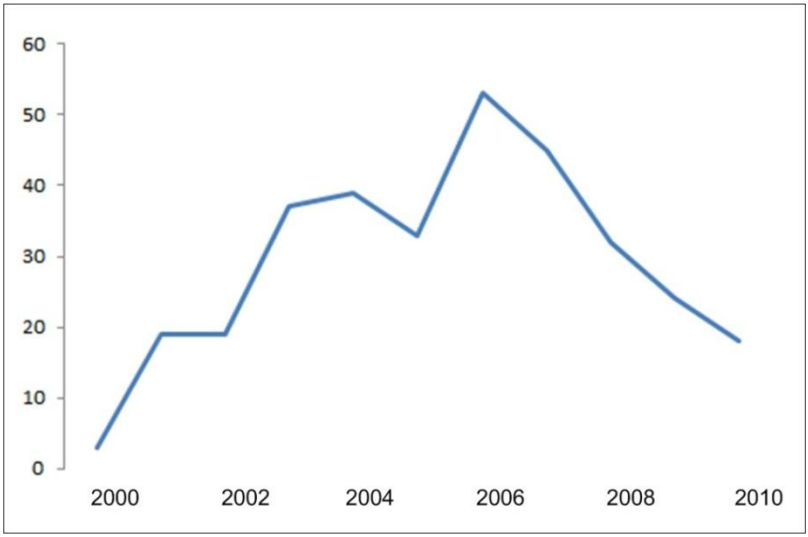


Figura 45 – Aderência entre os perfis de “Pacheco” e “Kern” no tempo (tema *Geral*).

Considerando que a associação entre todos os pares de conceitos *Pessoa* já está calculada no repositório, pode-se utilizar o serviço de *Redes de Relacionamentos* para projetar graficamente as redes formadas entre pessoas.

## 5.3 SERVIÇO *REDES DE RELACIONAMENTOS*

Os conceitos do tipo *Pessoa* não estão ligados diretamente entre si, assim, para estabelecermos as relações entre eles utiliza-se o peso de associação. Nesse caso, a associação representa a força do relacionamento baseado na comparação dos seus perfis. Assim, pesquisadores que possuem termos em comum tendem a ter uma relação mais forte na rede.

O serviço *Redes de Relacionamentos* implementado recebe como entrada o conjunto de conceitos do qual se deseja projetar a rede. O serviço identifica no repositório os relacionamentos de associação entre os conceitos do conjunto e gera o resultado de saída em formato

GraphML[16]. A rede representada em formato GraphML pode ser apresentada em qualquer software de visualização de redes que aceite esse formato. No exemplo ilustrado abaixo se utilizou o ISLinks®[17].

Tomando-se como exemplo o conjunto de 10 pessoas mais fortemente conectados ao conceito "Gestão do Conhecimento" apresentado na Figura 43, tem-se a rede mostrada na Figura 46. Essa rede apresenta as 10 pessoas e todos os relacionamentos.

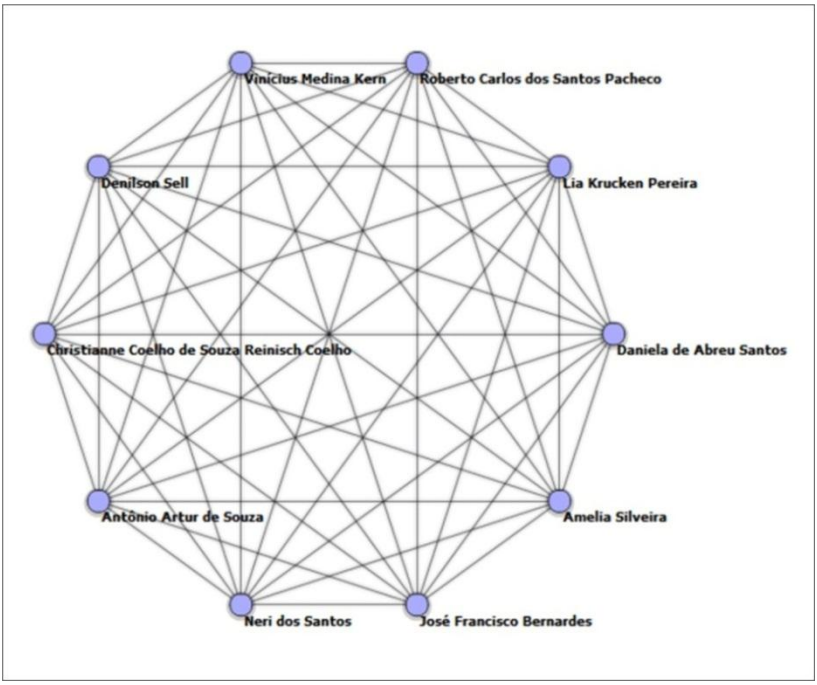


Figura 46 – Rede das pessoas mais fortemente conectados ao conceito "Gestão do Conhecimento" (tema *Geral*).

16 O GraphML (http://graphml.graphdrawing.org/) é um formato de arquivo utilizado para representar grafos e redes.

17 O ISLinks® é um componente da Plataforma ISEKP (*Enterprise Knowledge Platform*), do Instituto Stela (www.stela.org.br), para visualização de grafos e redes.

O ISLinks® permite ao usuário interagir com a aplicação por meio de um componente de *slider* para destacar as relações mais fortes. Veja a Figura 47.

Outra possibilidade é visualizar apenas os relacionamentos de um dado conceito. Por exemplo, a Figura 48 mostra as 20 pessoas mais fortemente relacionados ao conceito "Pacheco". O usuário pode, por exemplo, clicar sobre a aresta que liga "Pacheco" a "Sell" e o sistema apresenta a lista de conceitos (palavras-chave) comuns aos dois pesquisadores e que foram utilizados para geração da força de associação.

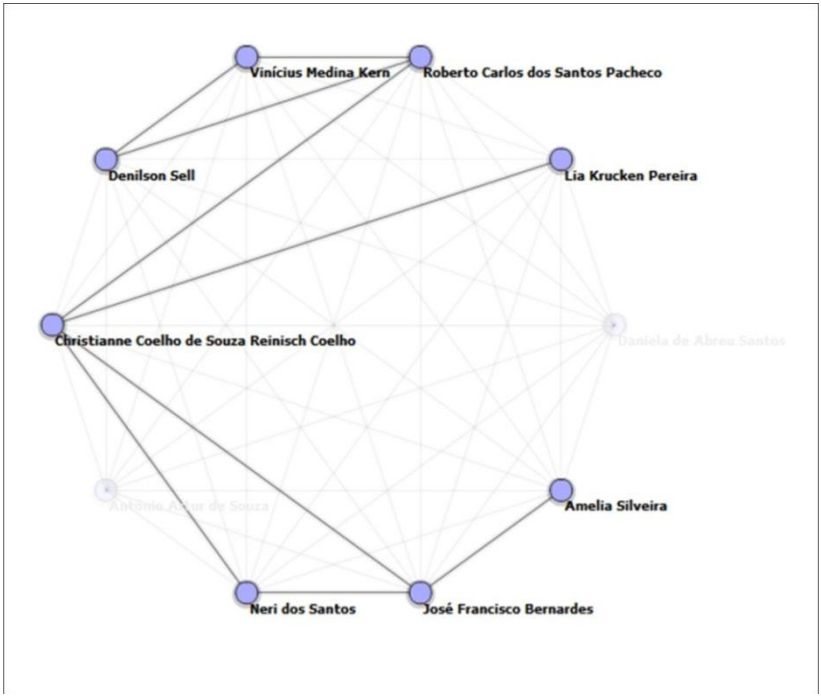


Figura 47 – Rede de pessoas ligadas a "Gestão do Conhecimento" com um corte (tema *Geral*).

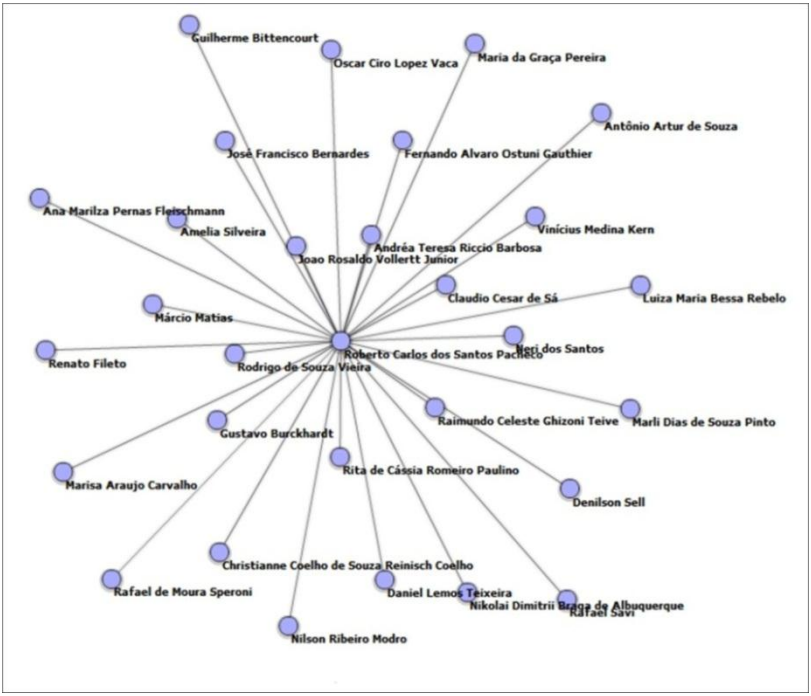


Figura 48 – Rede com os 20 pesquisadores mais fortemente conectados a "Pacheco" (tema *Geral*).

Assim como os perfis, as redes podem ser visualizadas considerando-se a dimensão tempo. Por exemplo, a Figura 49 mostra as 20 pessoas mais fortemente relacionados ao conceito "Pacheco" nos anos de 2003, 2004, 2005 e 2006. É possível ver as mudanças nas conexões de "Pacheco" ao longo do tempo. Por exemplo, percebe-se que "Nunes" está na rede em 2003 e não aparece mais nas redes dos 3 anos seguintes. Já "Fileto", que estava na rede em 2003, desaparece em 2004 e 2005 e volta em 2006.

A maioria dos conceitos estão presentes nas 4 redes (2003, 2004, 2005 e 2006), mas com pesos diferentes em cada ano. Esse é o caso, por exemplo, de "Sell", que possui o peso 0,334 em 2003, 0,191 em 2004, 0,096 em 2005 e 0,255 em 2006. Não é possível ver os pesos nas arestas (ligações) entre os nodos (pessoas) devido a uma limitação desta versão do componente ISLinks®.


Figura 49 – Redes de “Pacheco” por ano (tema *Geral*).

As redes aqui mostradas são todas pertencentes ao contexto/tema *Geral*. Contudo, assim como para os perfis, é possível ver as redes para cada um dos outros 5 temas/contextos (*Pesquisador*, *Docente*, *CompetenciaInovacao, Extensionista e Gestor*).

## 5.4 COMPARAÇÃO DO MODELO PROPOSTO COM OUTROS MODELOS DE KDT

Pretende-se analisar as contribuições do modelo proposto à área de descoberta de conhecimento em textos por meio uma análise comparativa do modelo proposto com outros dois modelos de KDT (MOONEY; NAHM, 2005; GONÇALVES, 2006).

### 5.4.1 Modelo proposto e o modelo de Mooney e Nahm (2005).

A principal novidade no trabalho de Mooney e Nahm (2005) está na proposição do modelo de KDT, para dados não estruturados, baseado no KDD, que lida com dados estruturados. Visão está que também foi seguida no modelo proposto. Nesse sentido, tanto o modelo de Mooney e Nahm (2005) (Figura 50, à esquerda), como o modelo proposto (Figura 50, à direita) possuem três etapas principais: (a) pré-processamento, (b) mineração de textos e (c) pós-processamento. Os dois modelos também são incrementais e iterativos com a participação dos usuários na etapa de interpretação e avaliação dos resultados.

As semelhanças entre os dois modelos está restrita ao parágrafo anterior. Pois, o modelo de Mooney e Nahm (2005) utiliza algumas técnicas de extração de informação para obter elementos textuais e apresenta um módulo de mineração para descoberta de regras de associação (sem utilizar o tempo).

Assim, todos os demais elementos presentes no modelo proposto (Capítulo 0) e do protótipo (Capítulo 4) não podem ser comparados. Como já foi apresentado, apenas a visão macro do processo de KDT apresentada por Mooney e Nahm (2005) está relacionada com o modelo proposto.

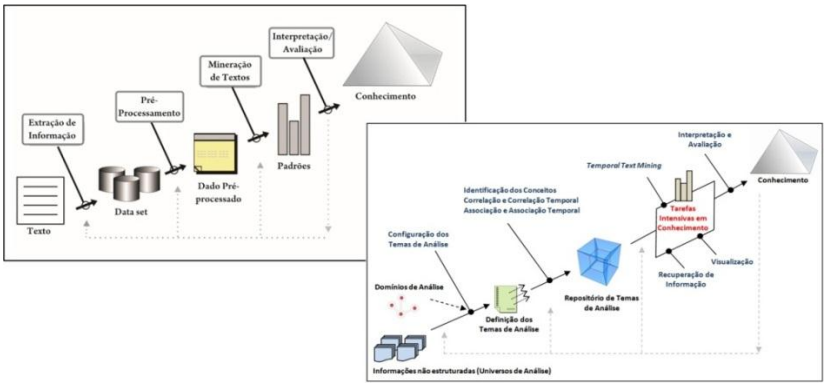


Figura 50 – Modelo de Mooney (MOONEY; NAHM, 2005) (à esquerda) e o modelo proposto (à direita).

### 5.4.2 Modelo proposto e o modelo de Gonçalves (2006).

Pretende-se analisar as contribuições do modelo proposto à área de descoberta de conhecimento em textos por meio uma análise comparativa com o modelo proposto por Gonçalves (2006).

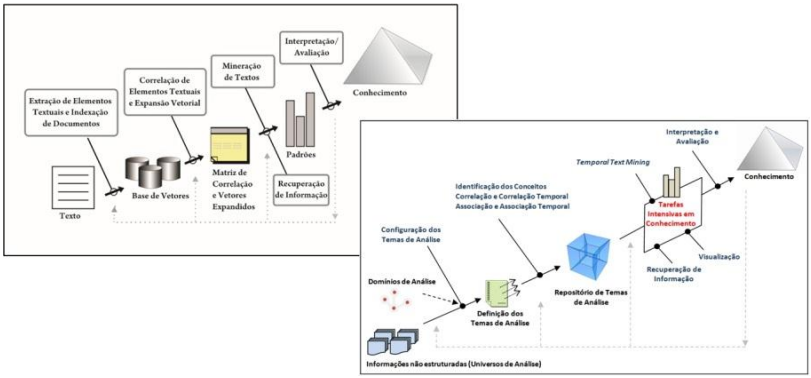


Figura 51 – Modelo de Gonçalves (GONÇALVES, 2006) (à esquerda) e o modelo proposto (à direita).

Os dois modelos (Figura 51) possuem três etapas principais: (a) pré-processamento, (b) mineração de textos e (c) e pós-processamento. Os dois modelos também são incrementais e iterativos com a participação dos usuários na etapa de interpretação e avaliação dos resultados. A principal diferença do modelo proposto está na inclusão da dimensão tempo no processo de descoberta de conhecimento em textos. Abaixo se discute alguns aspectos das duas propostas, envolvendo itens do modelo propriamente dito e de suas respectivas implementações.

a) Temas de Análise

O modelo de Gonçalves (2006) emprega como insumo algum tipo de fonte de informação não estruturada e utiliza implicitamente alguma forma de conhecimento de domínio (ex.: tabelas léxicas) no processo de extração de entidades.

O modelo proposto prevê explicitamente o uso de conhecimento de domínio nas análises e generaliza a combinação “fonte de

informação-conhecimento de domínio" por meio do conceito *temas de análise*. Um tema de análise consiste em um universo de análise e um domínio de análise. O universo de análise corresponde às fontes de informação a serem analisadas. O domínio de análise corresponde ao conhecimento de domínio utilizado, que pode estar representado por uma ontologia, tesauro, dicionário, vocabulário, etc.

b) Identificação das Ocorrências dos Conceitos

Tanto o modelo proposto nesta tese como o modelo proposto por Gonçalves (2006) necessitam de uma fase de identificação das ocorrências dos conceitos nos documentos textuais utilizando métodos de EI. A principal diferença do modelo proposto está na associação de um *timestamp* às ocorrências identificadas em cada documento por meio de metadados, extração de entidades ou *snapshots*.

c) Correlação

A correlação de elementos textuais é a parte principal do modelo proposto por Gonçalves (2006). Além de utilizar os métodos de correlação existente na literatura, esse modelo ainda apresenta um novo método, o LRD. Esse método utiliza, além das coocorrências, a distância entre os elementos textuais.

O modelo proposto nesta tese apenas utiliza os métodos baseados em coocorrência disponíveis na literatura. A novidade aqui está na possibilidade de se criar análises que utilizem diferentes métodos de correlação e também diferentes níveis de coocorrência (sentença, janela, documento, etc.). Cada uma destas relações podem ser modeladas utilizando diferentes tipos de relações (classe *Relation*).

d) Correlação Temporal

Uma das principais diferenças do modelo proposto em relação ao modelo de Gonçalves (2006) é a incorporação da dimensão tempo nas relações. Assim, além da fase de correlação (na qual são calculados

pesos de correlação sem considerar o tempo), têm-se a fase de correlação temporal.

Na validação orientada à tarefa do método LRD realizada por Gonçalves (2006), o autor cita que alguns usuários tiveram dificuldade em ordenar "entidades relacionadas" do tipo *Pessoa*. Para isso, o autor sugeriu "que as relações pudessem considerar a questão temporal, visto que relações mais atuais tendem a ser mais relevantes". Isso aponta a necessidade de incluir a dimensão tempo na matriz de correlação.

e) Associação

O modelo de Gonçalves (2006) apresenta uma abordagem para identificação de relacionamentos indiretos entre conceitos baseada na expansão do espaço vetorial e em técnicas de agrupamento.

No modelo proposto nesta tese, os relacionamentos indiretos são tratados de maneira mais ampla por meio da associação. A matriz de associação apresentada no modelo possibilita a aplicação de conceitos da área de DBL, como a descoberta ABC.

f) Associação Temporal

A abordagem para identificação de relacionamentos indiretos entre conceitos, baseada na expansão do espaço vetorial e em técnicas de agrupamento, proposta por Gonçalves (2006) não lida com a questão temporal dos relacionamentos.

g) Repositório de Temas de Análise

O modelo de Gonçalves (2006) não apresenta estrutura semelhante ao repositório de temas de análise no qual as dimensões conceito (*source* e *target*), tempo, relação e tema são representadas. Além das informações sobre as matrizes de correlação, correlação temporal, associação e associação temporal, o repositório de temas de análise também armazena outras informações que podem ser úteis para as tarefas intensivas em conhecimento, tais como a frequência conjunta e individual dos conceitos.

h) Tarefas Intensivas em Conhecimento

O modelo proposto apresenta uma fase sobre a execução de tarefas intensivas em conhecimento (com ênfase nos aspectos temporais dos relacionamentos diretos e indiretos entre os conceitos do domínio). Nesta fase, os métodos de TTM e de visualização de informações temporais, auxiliados pela área de RI, são combinados para apoiar os usuários em diversas tarefas intensivas em conhecimento.

O modelo proposto por Gonçalves (2006) apresenta as fases (a) "Geração de Padrões", que tem por objetivo a identificação de relacionamentos indiretos entre entidades; e (b) "Visualização de Padrões", que é usado para a apresentação gráfica das conexões entre as entidades para facilitar o processo de descoberta de conhecimento.

i) Correlação Padrão

Este módulo (apresentado no Capítulo 4) realiza os passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. O modelo de Gonçalves (2006) apresenta um artefato semelhante. Porém, há duas diferenças principais: (a) no modelo de Gonçalves esse sistema não gera a matriz de Correlação Temporal, já que o modelo não incorpora a dimensão tempo, com citado anteriormente; e (b) no trabalho de Gonçalves, a correlação padrão faz parte do modelo, enquanto que no modelo proposto nesta tese esse módulo faz parte do protótipo baseado no modelo.

j) Correlação Rápida

Este módulo (apresentado no Capítulo 4) também realiza os passos de Identificação das Ocorrências dos Conceitos, Correlação e Correlação Temporal. Essa abordagem mostra-se útil em situações onde o número de instâncias do domínio de análise é pequeno e a quantidade de documentos do universo de análise é grande. O modelo de Gonçalves (2006) não apresenta essa abordagem.

## 5.5 CONSIDERAÇÕES FINAIS

Este capítulo está dividido em duas partes. A primeira descreveu um estudo de caso para demonstração de viabilidade do modelo proposto. O cenário de aplicação foi apresentado bem como os serviços de conhecimento utilizados: *Perfil de Conceitos* e *Redes de Relacionamentos*. A segunda parte do capítulo descreveu uma análise comparativa do modelo proposto com outros dois modelos de KDT (MOONEY; NAHM, 2005; GONÇALVES, 2006), com o objetivo de se discutir as contribuições do modelo proposto à área de descoberta de conhecimento em textos.

# 6 CONCLUSÕES E TRABALHOS FUTUROS

O objetivo geral desta tese é desenvolver um modelo de descoberta de conhecimento a partir de informações não estruturadas que possibilite analisar a evolução dos relacionamentos entre os elementos textuais ao longo do tempo. Para isso, foi proposto um modelo de *Temporal Knowledge Discovery in Texts*, baseado no modelo de KDT (etapa de pré-processamento, etapa de mineração de textos e etapa de pós-processamento), com ênfase no aspecto temporal dos relacionamentos entre os elementos textuais. Trata-se de um modelo que estende os modelos de KDT de Mooney e Nahm (2005) e de Gonçalves (2006), acrescentando novos elementos, sendo a mais importante a dimensão temporal nos relacionamentos entre os conceitos do domínio.

O modelo proposto é dividido por fases, assim como os modelos tradicionais de descoberta de conhecimento. As fases deste modelo são: configuração dos temas de análise, identificação das ocorrências dos conceitos, correlação e correlação temporal, associação e associação temporal, criação do repositório de temas de análise, e tarefas intensivas em conhecimento, com ênfase nos relacionamentos diretos e indiretos entre os conceitos do domínio. Cada uma destas fases foi explicada em detalhes utilizando-se como exemplo uma fonte de informação não estruturada e temporal e um conjunto de instâncias de uma ontologia como domínio de análise. Na fase de tarefas intensivas em conhecimento, as tarefas de *geração de vetores de contexto* e *descoberta abc* foram examinadas em detalhes.

Enquanto modelo de KDT, além da incorporação da dimensão tempo, o TKDT permite criação de temas de análise que propiciam flexibilidade na combinação de diferentes fontes de informação com diferentes formas de conhecimento de domínio. Também apresenta uma visão integrada dos relacionamentos ao longo do tempo. E é voltado para aplicações Engenharia e Gestão do Conhecimento, pois possibilita a execução de tarefas intensivas em conhecimento.

Para atingir o objetivo geral, um dos objetivos específicos é investigar e propor uma forma de se identificar e representar o peso dos relacionamentos diretos e indiretos entre os elementos textuais ao longo do tempo. Para isso, pesquisaram-se as áreas de correlação e associação de elementos textuais, modelos baseados em coocorrência, modelo espaço vetorial, similaridade de vetores e estruturas de indexação.

Quanto aos relacionamentos diretos, foi proposta uma representação por meio das matrizes de correlação e correlação temporal. Logo, trata-se de um modelo que permite analisar os relacionamentos diretos que podem ser calculados utilizando diferentes níveis de coocorrência (ex.: sentença, parágrafo, janela e documento) e diferentes métodos baseados em coocorrência (ex.: *Chi-square* ($x^2$), *Z score*, *Phisquared* ($\Phi^2$), IM, etc.). Quanto aos relacionamentos indiretos, foi proposta uma representação por meio das matrizes de associação e associação temporal. Assim, trata-se de um modelo que permite analisar relacionamentos indiretos entre elementos textuais, por meio da aplicação de conceitos da área de DBL. Além disso, os pesos desses relacionamentos podem ser calculados utilizando diferentes medidas de similaridade.

Outro objetivo específico é identificar na literatura métodos, técnicas e algoritmos relativos a correlação, associação e análise temporal de informações textuais, que possam ser utilizados na etapa de TTM do modelo proposto. Assim, os métodos de TTM, em conjunto com a área de visualização de informações temporais e auxiliados pela área de RI, são combinados para apoiar os usuários em tarefas intensivas em conhecimento.

Foi também definido como objetivo específico a demonstração da viabilidade do modelo proposto por meio do desenvolvimento de um protótipo e sua aplicação em um estudo de caso. Esse protótipo possui um módulo que permite aos usuários a configuração de temas de análise; um módulo de indexação das fontes de informação; dois módulos de correlação, chamados de "Correlação Rápida" e "Correlação Padrão", que são responsáveis pelas fases de identificação das ocorrências dos conceitos, correlação e correlação temporal; um módulo para cálculo da força de associação entre os conceitos; e um módulo de serviços de conhecimento. As informações do repositório de temas de análise são mapeadas em um modelo de dados dimensional. Já o estudo de caso foi realizado com base de 1.000 currículos, em formato XML, de pesquisadores da UFSC. Os serviços *Perfil de Conceitos* e *Redes de Relacionamentos* foram utilizados sobre as palavras-chave de itens relativos à produção científica, formação acadêmica e atividade profissional de cada pessoa.

O último objetivo específico consiste em analisar as contribuições do modelo proposto à área de descoberta de conhecimento em textos por meio de uma análise comparativa com outros modelos existentes na

literatura. Para isso, foi realizada uma análise comparativa do modelo proposto com outros dois modelos de KDT (MOONEY; NAHM, 2005; GONÇALVES, 2006), a qual destacou as contribuições do modelo proposto à área de descoberta de conhecimento em textos.

## 6.1 TRABALHOS FUTUROS

Como apresentado anteriormente, o domínio de análise refere-se ao conhecimento de domínio utilizado nas análises. O domínio de análise é formado por um conjunto de instâncias da área de interesse. Esse conhecimento de domínio pode estar representado em ontologias, tesauros, taxonomias, dicionário, vocabulários, etc. Para facilitar a aquisição desse conhecimento, sugere-se pesquisar como métodos para a manutenção e/ou população de ontologias de maneira automática ou semiautomática (CIMIANO; VOLKER, 2005; WEGRZYN-WOLSKA; SZCZEPANIAK et al., 2007; FORTUNA; LAVRAČ et al., 2008; GACITUA; SAWYER et al., 2008; CECI; SILVA et al., 2010) podem ser integrados ao modelo proposto, na etapa de Configuração de Temas de Análise, para auxiliar na definição do domínio de análise.

Outra sugestão de trabalho futuro é a pesquisa na área de extração de expressões temporais (ALONSO; GERTZ et al., 2009; STROTGEN; GERTZ et al., 2010). Como exposto anteriormente, a forma e a possibilidade de se obter uma marca temporal para cada documento dependem das características da cada fonte de informação. Contudo, a inclusão no modelo de métodos para extração de expressões temporais, na fase de Identificação das Ocorrências dos Conceitos, pode ser interessante em determinados domínios de aplicação.

Uma possibilidade de trabalho de futuro refere-se à investigação de possíveis formas de se representar o Repositório de Temas de Análise, apresentado conceitualmente pela ontologia mostrada na Figura 22. Nessa tese, foi proposta uma representação por meio de um modelo de dados dimensional (Figura 33). Contudo, outras formas de representação podem ser utilizadas dependendo do caso concreto. Por exemplo, pode-se utilizar o *Bigtable* (CHANG; DEAN et al., 2006), que é um sistema de armazenagem distribuída para gerenciar grandes quantidades de dados estruturados. Assim, é possível ter uma representação das dimensões do repositório de temas de análise

distribuída entre diversos servidores e que pode crescer até *pentabytes* de dados.

Outra possibilidade de trabalho é a integração do protótipo desenvolvido com novas formas e ferramentas de visualização de dados. Um caminho a seguir é a revisão da literatura sobre abordagens visuais para informações textuais e temporais é apresentada por Šilić e Dalbelo Bašić (2010). Esse trabalho apresenta áreas relacionadas, tipos de coleções de dados que são visualizados, aspectos técnicos de geração de visualizações e metodologias de avaliação.

Por último, sugere-se também como trabalho futuro a pesquisa e o desenvolvimento de novos métodos, técnicas, algoritmos, etc., que possam ser utilizados em tarefas intensivas em conhecimento, que explorem novos aspectos da dimensão tempo nos relacionamentos diretos e indiretos entre os conceitos do domínio.

## REFERÊNCIAS BIBLIOGRÁFICAS

ABE, H.; TSUMOTO, S. **Detecting temporal patterns of technical phrases by using importance indices in a research documents**. Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics. San Antonio, TX, USA: IEEE Press 2009.

AHMAD, K.; AL-THUBAITY, A. **Can text analysis tell us something about technology progress?** Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20. Sapporo, Japan: Association for Computational Linguistics 2003.

ALAKO, B. et al. CoPub Mapper: mining MEDLINE based on search term co-publication. **BMC Bioinformatics,** v. 6, n. 1, p. 51, 2005. ISSN 1471-2105.

ALLAN, J. Introduction to Topic Detection and Tracking. In: (Ed.). **Topic Detection and Tracking: Event-Based Information Organization**: Kluwer Academic Publishers, 2002. p.1-16. ISBN 0-7923-7664-1.

ALLAN, J.; PAPKA, R.; LAVRENKO, V. **On-line new event detection and tracking**. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne, Australia: ACM 1998.

ALONSO, O.; GERTZ, M.; BAEZA-YATES, R. **Clustering and exploring search results using timeline constructions**. Proceeding of the 18th ACM conference on Information and knowledge management. Hong Kong, China: ACM 2009.

ARNING, A.; RAGHAVAN, R. A. P. **A Linear Method for Deviation Detection in Large Databases**. International Conference on Knowledge Discovery and Data Mining. Portland (USA)**:** 164-169 p. 1996.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. ACM Press, Addison-Wesley, 1999.

BAHARUDIN, B.; LEE, L. H.; KHAN, K. A Review of Machine Learning Algorithms for Text-Documents Classification. **Journal of Advances in Information Technology,** v. 1, n. 1, p. 4-20, 2010. ISSN 17982340.

BAKER, N. C.; HEMMINGER, B. M. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. **Journal of Biomedical Informatics,** v. 43, n. 4, p. 510-519, 2010. ISSN 1532-0464.

BALANCIERI, R. et al. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. **Ciência da Informação,** v. 34, n. 1, p. 64-77, 2005.

BARABÁSI, A.-L. **Linked: how everything is connected to everything else and what it means for business, science, and everyday**. New York: Plume, 2003.

BARBOSA-SILVA, A. et al. LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. **BMC Bioinformatics,** v. 11, n. 1, 2010. ISSN 1471-2105.

BERENDT, B.; SUBASIC, I. Measuring graph topology for interactive temporal event detection. **Künstliche Intelligenz,** v. 2, p. 11-17, 2009. ISSN 0933-1875.

BERRY, M. J. A.; LINOFF, G. **Data mining techniques - for marketing, sales, and customer support**. New York: John Wiley & Sons, 1997.

BÖTTCHER, M.; HÖPPNER, F.; SPILIOPOULOU, M. On exploiting the power of time in data mining. **SIGKDD Explor. Newsl.,** v. 10, n. 2, p. 3-11, 2008. ISSN 1931-0145.

BOUANDAS, K.; OSMANI, A. **Mining Association Rules in Temporal Sequences**. Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007). Honolulu, HI, **:** 610-615 p. 2007.

BROWNE, F. et al. From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions. **Advances in Artificial Intelligence,** v. 2010, n. ID 924529, p. 15, 2010. ISSN 1687-7470.

BUI, Q.-C. et al. Extracting causal relations on HIV drug resistance from literature. **BMC Bioinformatics,** v. 11, n. 1, p. 101, 2010. ISSN 1471-2105.

CECI, F. et al. **Towards a Semi-Automatic Approach for Ontology Maintenance**. 7th CONTECSI International Conference on Information Systems and Technology Management. São Paulo (SP) 2010.

CHANG, F. et al. **Bigtable: A Distributed Storage System for Structured Data**. OSDI'06: Seventh Symposium on Operating System Design and Implementation. Seattle, WA 2006.

CHANG, J. T.; ALTMAN, R. B. Extracting and characterizing gene-drug relationships from the literature. **Pharmacogenetics and Genomics,** v. 14, n. 9, p. 577-586, 2004. ISSN 1744-6872.

CHEN, H.; SHARP, B. Content-rich biological network constructed by mining PubMed abstracts. **BMC Bioinformatics,** v. 5, n. 1, p. 147, 2004. ISSN 1471-2105.

CHEN, L.-C. Using a two-stage technique to design a keyword suggestion system. **Information Research,** v. 15, n. 1, 2010. ISSN 1368-1613.

CHEN, W. et al. Online detection of bursty events and their evolution in news streams. **Journal of Zhejiang University - Science C,** v. 11, n. 5, p. 340-355, 2010.

CHEUNG, W. M. Ontological approach of organizational knowledge to support collaborative product development. **Journal of Advanced Manufacturing Systems,** v. 5, n. 1, p. 3-25, 2006.

CHURCH, K. W.; GALE, W. A. **Concordances for Parallel Text**. Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research. Oxford, England**:** 40-62 p. 1991.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics,** v. 16, n. 1, p. 22-29, 1990. ISSN 0891-2017.

CIMIANO, P.; VOLKER, J. **Text2Onto - A Framework for Ontology Learning and Datadriven Change Discovery** 2005.

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, January 1, 2005 2005.

COHEN, T. Exploring MEDLINE space with random indexing and pathfinder networks. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 126-130, 2008. ISSN 1942-597X.

COHEN, T.; SCHVANEVELDT, R.; WIDDOWS, D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. **Journal of Biomedical Informatics,** v. 43, n. 2, p. 240-256, 2010. ISSN 1532-0464.

COHEN, T.; SCHVANEVELDT, R. W.; RINDFLESCH, T. C. **Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space**. AMIA Annu Symp Proc. 2009**:** 114–118 p. 2009.

CONRAD, J. G.; UTT, M. H. **A system for discovering relationships by feature extraction from text databases**. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland: Springer-Verlag New York, Inc. 1994.

CORY, K. Discovering Hidden Analogies in an Online Humanities Database. **Computers and the Humanities,** v. 31, n. 1, p. 1-12, 1997.

DAI, H.-J. et al. New Challenges for Biological Text-Mining in the Next Decade. **Journal of Computer Science and Technology,** v. 25, n. 1, p. 169-179, 2010. ISSN 1000-9000.

DAVIDOV, D.; RAPPOPORT, A. **Classification of Semantic Relationships between Nominals Using Pattern Clusters**. Proceedings of ACL'08**:** 227-235 p. 2008.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American Society for Information Science,** v. 41, p. 391-407, 1990.

DENG, Q.; YU, D. Mapping Knowledge in Product Development through Process Modelling. **Journal of Information & Knowledge Management (JIKM),** v. 5, n. 03, p. 233-242, 2006.

DING, C. H. Q. **A Probabilistic Model for Dimensionality Reduction in Information Retrieval and Filtering**. In Proc. of 1st SIAM Computational Information Retrieval Workshop. Raleigh, NC 2000.

DÖRRE, J.; GERSTL, P.; SEIFFERT, R. **Text mining: finding nuggets in mountains of textual data**. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States: ACM 1999.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing and Management: an International Journal,** v. 38, n. 6, p. 823-848, 2002. ISSN 0306-4573.

EOGHAN, D. H.; LARS, J. J.; PEER, B. Predicting biological networks from genomic data. **FEBS letters,** v. 582, n. 8, p. 1251-1258, 2008. ISSN 0014-5793.

ERHARDT, R.; SCHNEIDER, R.; BLASCHKE, C. Status of text-mining techniques applied to biomedical text. **Drug Discovery Today,** v. 11, p. 315 - 325, 2006.

FAYYAD, U. M. Data Mining and Knowledge Discovery: Making Sense Out of Data. **IEEE Expert: Intelligent Systems and Their Applications,** v. 11, n. 5, p. 20-25, 1996. ISSN 0885-9000.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: (Ed.). **Advances in knowledge discovery and data mining**: American Association for Artificial Intelligence, 1996a. p.1-34. ISBN 0-262-56097-6.

_____. Knowledge discovery and data mining: Towards a unifying framework. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996b. Portland, Oregon. AAAI Press. p.82-88.

FELDMAN, R. et al. **Trend Graphs: Visualizing the Evolution of Concept Relationships in Large Document Collections**. Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery: Springer-Verlag 1998.

FELDMAN, R.; DAGAN, I. **Knowledge Discovery in Textual Databases (KDT)**. Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95) 1995.

FELDMAN, R.; DAGAN, I.; HIRSH, H. Mining Text Using Keyword Distributions. **Journal of Intelligent Information Systems,** v. 10, n. 3, p. 281-300, 1998. ISSN 0925-9902.

FELDMAN, R. et al. Knowledge Management: A Text Mining Approach. Proc. the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98), 1998. 1998. p.9.1-9.10.

FORTUNA, B.; LAVRAČ, N.; VELARDI, P. Advancing Topic Ontology Learning through Term Extraction. In: (Ed.). **PRICAI 2008: Trends in Artificial Intelligence**: Springer Berlin / Heidelberg, v.5351, 2008. p.626-635. (Lecture Notes in Computer Science).

FUNG, G. P. C. et al. **Parameter free bursty events detection in text streams**. Proceedings of the 31st international conference on Very large data bases. Trondheim, Norway: VLDB Endowment 2005.

GACITUA, R.; SAWYER, P.; RAYSON, P. A flexible framework to experiment with ontology learning techniques. **Know.-Based Syst.,** v. 21, n. 3, p. 192-199, 2008. ISSN 0950-7051.

GANDRA, P.; PRADHAN, M.; PALAKAL, M. J. Identification of biological relationships from text documents using efficient computational methods. **Journal of Bioinformatics and Computational Biology (JBCB)** v. 1, n. 2, p. 307-342, 2003.

_____. **Biomedical association mining and validation**. Proceedings of the International Symposium on Biocomputing. Calicut, Kerala, India: ACM 2010.

GANIZ, M. C.; POTTENGER, W. M.; JANNECK, C. D. Recent Advances in Literature Based Discovery. **Journal of the American Society for Information Science and Technology, JASIST**, 2006.

GARTEN, Y.; ALTMAN, R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. **BMC Bioinformatics,** v. 10, n. Suppl 2, p. S6, 2009. ISSN 1471-2105.

GARTEN, Y.; TATONETTI, N. P.; ALTMAN, R. B. **Improving the Prediction of Pharmacogenes Using Using Text-Derived Drug-Gene Relationships**. Proceedings of the Pacific Symposium Biocomputing. Kamuela, Hawaii, USA 2010.

GHARIB, T. F. et al. An efficient algorithm for incremental mining of temporal association rules. **Data & Knowledge Engineering,** v. 69, n. 8, p. 800-815, 2010. ISSN 0169-023X.

GIOVINAZZO, W. A. **Object-Oriented Data Warehouse Design - Building a Star Schema**. New Jersey: Prentice Hall, 2000.

GONÇALVES, A. et al. **LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval**. Advances in Web-Age Information Management, 7th International Conference (WAIM 2006). Hong Kong, China**:** 122-133 p. 2006.

GONÇALVES, A. L. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento**. 2006. 196 (Doutorado). Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis (SC).

GONÇALVES, A. L. et al. **A Text Mining Approach towards Knowledge Management Applications**. Proceedings of the International Workshop on Information Retrieval on Current Research Information Systems. Copenhagen, Denmark**:** 7-28 p. 2006.

GOORHA, S.; UNGAR, L. **Discovery of significant emerging trends**. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Washington, DC, USA: ACM 2010.

GORDON, M.; LINDSAY, R. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. **Journal of the American Society for Information Science,** v. 47, n. 2, p. 116-128, 1996. ISSN 0002-8231.

GORDON, M.; LINDSAY, R. K.; FAN, W. Literature-based discovery on the World Wide Web. **ACM Transactions on Internet Technology (TOIT),** v. 2, n. 4, p. 261-275, 2002. ISSN 1533-5399.

GORDON, M. D.; DUMAIS, S. Using latent semantic indexing for literature based discovery. **J. Am. Soc. Inf. Sci.,** v. 49, n. 8, p. 674-685, 1998. ISSN 0002-8231.

GREENGRASS, E. **Information Retrieval: A Survey**. 2000. 224

GUPTA, V.; LEHAL, G. S. A Survey of Text Mining Techniques and Applications. **Journal of Emerging Technologies in Web Intelligence,** v. 1, n. 1, p. 60-76, 2009. ISSN 1798-0461.

HA-THUC, V. et al. **Event Intensity Tracking in Weblog Collections**. Proceedings of the 3rd International AAAI Conference on

Weblogs and Social Media Data Challenge Workshop. San Jose, California, USA 2009.

_____. **A relevance-based topic model for news event tracking**. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, MA, USA: ACM 2009.

HAIR, J. F. et al. **Multivariate data analysis**. 5th. Prentice Hall; 5th edition (March 23, 1998), 1998. 768

HAVRE, S. et al. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. **IEEE Transactions on Visualization and Computer Graphics,** v. 8, n. 1, p. 9-20, 2002. ISSN 1077-2626.

HE, Q. et al. **Bursty Feature Representation for Clustering Text Streams**. SIAM International Conference on Data Mining (SDM) 2007.

_____. **Detecting topic evolution in scientific literature: how can citations help?** Proceeding of the 18th ACM conference on Information and knowledge management. Hong Kong, China: ACM 2009.

HE, R. et al. Cascaded Regression Analysis Based Temporal Multi-document Summarization. **Informatica - An International Journal of Computing and Informatics,** v. 34, n. 1, 2010. ISSN 1854-3871.

HENDRIKS, P. H. J. Do smarter systems make for smarter organizations? **Decision Support Systems,** v. 27, n. 1-2, p. 197-211, 1999. ISSN 0167-9236.

HIMMA, K. The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition. **Ethics and Information Technology,** v. 9, n. 4, p. 259-272, 2007.

HOFFMANN, R.; VALENCIA, A. Implementing the iHOP concept for navigation of biomedical literature. **Bioinformatics,** v. 21, n. suppl_2, p. ii252-258, September 1, 2005 2005.

HOLSAPPLE, W. The inseparability of modern knowledge management and computer-based technology. **Journal of Knowledge Management,** v. 9, n. 1, p. 42-52, 2005. ISSN 1367-3270.

HOLZ, F.; TERESNIAK, S. **Towards Automatic Detection and Tracking of Topic Change**. Computational Linguistics and Intelligent Text Processing: Springer Berlin / Heidelberg. 6008**:** 327-339 p. 2010.

HRISTOVSKI, D. et al. Using literature-based discovery to identify disease candidate genes. **International Journal of Medical Informatics,** v. 74, n. 2-4, p. 289-298, 2005. ISSN 1386-5056.

HUANG, H.-C. Designing a knowledge-based system for strategic planning: A balanced scorecard perspective. **Expert Syst. Appl.,** v. 36, n. 1, p. 209-218, 2009. ISSN 0957-4174.

INMON, W. H. **Como construir o Data Warehouse**. Rio de Janeiro: Campus, 1997.

JONES, W. P.; FURNAS, G. W. Pictures of relevance: a geometric analysis of similarity measures. **Journal of the American Society for Information Science,** v. 38, n. 6, p. 420-442, 1987. ISSN 0002-8231.

JORGE-BOTANA, G. et al. **Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus**. 2009. 424-40 ISBN 1138-7416.

JUSTESON, J. S.; KATZ, S. M. Technical terminology: some linguistic properties and an algorithm for identification in text. **Natural Language Engineering,** v. 1, n. 01, p. 9-27, 1995. ISSN 1351-3249.

KAMARUDDIN, S. S.; HAMDAN, A. R.; BAKAR, A. A. **Text Mining for Deviation Detection in Financial Statement**. Proceedings of the International Conference on Electrical Engineering and Informatics. Bandung, Indonesia 2007.

KASTRIN, A.; HRISTOVSKI, D. A fast document classification algorithm for gene symbol disambiguation in the BITOLA literature-based discovery support system. **AMIA Annual Symposium**

**proceedings AMIA Symposium AMIA Symposium (2008)**, p. 358-362, 2008.

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. ISSN 1386-145X.

KIM, Y. et al. **Automatic discovery of technology trends from patent text**. Proceedings of the 2009 ACM symposium on Applied Computing. Honolulu, Hawaii: ACM 2009.

KIM, Y. G.; SUH, J. H.; PARK, S. C. Visualization of patent analysis for emerging technology. **Expert Systems with Applications: An International Journal,** v. 34, n. 3, p. 1804-1812, 2008. ISSN 0957-4174.

KIMBALL, R. et al. **The Data Warehouse lifecycle toolkit: expert methods for designing, developing and deploying Data Warehouses**. New York: John Wiley & Sons, 1998.

KIMBALL, R.; ROSS, M. **Data Warehouse toolkit: the complete guide to dimensional modeling**. New York: John Wiley & Sons, 2002.

KLEINBERG, J. **Bursty and hierarchical structure in streams**. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada: ACM 2002.

KNORR, E. M.; NG, R. T.; TUCAKOV, V. Distance-Based Outliers: Algorithms and Applications. **The VLDB Journal - The International Journal on Very Large Data Bases,** v. 8, n. 3-4, p. 237-253 2000. ISSN 1066-8888.

KOBAYASHI, M.; TAKEDA, K. Information retrieval on the web. **ACM Computing Surveys (CSUR),** v. 32, n. 2, p. 144-173, 2000. ISSN 0360-0300.

KOED. International Conference on Knowledge Engineering and Ontology Development. 2009. Disponível em: < http://www.keod.ic3k.org >.

KONTOSTATHIS, A. et al. **A Survey of Emerging Trend Detection in Textual Data Mining**. Springer, 2004.

KORFHAGE, R. R. **Information storage and retrieval.** New York: Wiley Computer Publishing, 1997.

KOWALSKI, G. **Information retrieval systems :theory and implementation**. Boston: Kluwer Academic Publishers, 1997. 300

LEE, C.-H.; LIN, C.-R.; CHEN, M.-S. **On Mining General Temporal Association Rules in a Publication Database**. Proceedings of the 2001 IEEE International Conference on Data Mining: IEEE Computer Society 2001.

LENT, B.; AGRAWAL, R.; SRIKANT, R. **Discovering Trends in Text Databases**. Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD: AAAI Press**:** 227-230 p. 1997.

LEVY, D. M. **To grow in wisdom: vannevar bush, information overload, and the life of leisure**. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. Denver, CO, USA: ACM 2005.

_____. **More, Faster, Better: Governance in an Age of Overload, Busyness, and Speed.** The Emergence of Governance in Global Cyberspace 2006.

LI, Z. et al. **A probabilistic model for retrospective news event detection**. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Salvador, Brazil: ACM 2005.

LIN, Y. R. et al. Analyzing communities and their evolutions in dynamic social networks. **ACM Trans. Knowl. Discov. Data,** v. 3, n. 2, p. 1-31, 2009. ISSN 1556-4681.

LINDSAY, R. K.; GORDON, M. D. Literature-based discovery by lexical statistics. **Journal of the American Society for Information Science and Technology,** v. 50, n. 7, p. 574-587, 1999. ISSN 0002-8231.

LIPNAK, J.; STAMP, J. **Networks, redes de conexão: pessoas conectando-se com pessoas**. São Paulo: Aquarela, 1992.

LIU, H.; FRIEDMAN, C. **Mining terminological knowledge in large biomedical corpora**. Pacific Symposium On Biocomputing Pacific Symposium On Biocomputing. 426**:** 415-426 p. 2003.

LIU, S. et al. **A sentence level probabilistic model for evolutionary theme pattern mining from news corpora**. Proceedings of the 2009 ACM symposium on Applied Computing. Honolulu, Hawaii: ACM 2009.

LYMAN, P. **How Much Information?** USA: University of California 2000.

_____. **How Much Information?** USA: University of California 2003.

MAKKONEN, J.; AHONEN-MYKA, H.; SALMENKIVI, M. Simple Semantics in Topic Detection and Tracking. **Information Retrieval,** v. 7, n. 3-4, p. 347-368, 2004. ISSN 1386-4564.

MANNILA, H. **Data Mining: Machine Learning, Statistics, and Databases**. Proceedings of the Eighth International Conference on Scientific and Statistical Database Management: IEEE Computer Society 1996.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, Massachusetts: The MIT Press, 1999.

MEI, Q. et al. **A probabilistic approach to spatiotemporal theme pattern mining on weblogs**. Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland: ACM 2006.

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

MENGLE, S. S. R.; GOHARIAN, N. **Mining temporal relationships among categories**. Proceedings of the 2010 ACM Symposium on Applied Computing. Sierre, Switzerland: ACM 2010.

MESQUITA, F.; MERHAV, Y.; BARBOSA, D. **Extracting Information Networks from the Blogosphere: State-of-the-Art and Challenges**. 4th Int'l AAAI Conference on Weblogs and Social Media--Data Challenge. Washington, DC 2010.

MITRA, M.; CHAUDHURI, B. B. Information Retrieval from Documents: A Survey. **Information Retrieval,** v. 2, n. 2-3, p. 141-163, 2000. ISSN 1386-4564.

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.

MOON, I.-C. et al. **Temporal Issue Trend Identifications in Blogs**. International Conference on Computational Science and Engineering. Vancouver, Canada. 4**:** 619-626 p. 2009.

MOONEY, R. J.; NAHM, U. Y. **Text Mining with Information Extraction**. Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium. DAELEMANS, W., DU PLESSIS, T., SNYMAN, C. AND TECK, L. Bloemfontein, South Africa: Van Schaik Pub.**:** 141-160 p. 2005.

MÖRCHEN, F. et al. **Anticipating annotations and emerging trends in biomedical literature**. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA: ACM 2008.

_____. **Emerging Trend Prediction in Biomedical Literature**. AMIA Annu Symp Proc. : American Medical Informatics Association 2008.

NASUKAWA, T.; NAGANO, T. Text analysis and knowledge mining system. **IBM Systems Journal,** v. 40, n. 4, p. 967-984, 2001. ISSN 0018-8670.

NEWMAN, M. E. J. **Clustering and Preferential Attachment in Growing Networks**. Santa Fe Institute. 2001

NISSEN, M. E. **Harnessing knowledge dynamics**. Hershey, PA: Idea Group Inc., 2006.

NIU, Y.; OTASEK, D.; JURISICA, I. Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I2D. **Bioinformatics,** v. 26, n. 1, p. 111-119, January 1, 2010 2010.

NØRVÅG, K.; ERIKSEN, T. Ø.; SKOGSTAD, K.-I. Mining Association Rules in Temporal Document Collections. In: (Ed.). **Foundations of Intelligent Systems**: Springer Berlin / Heidelberg, v.4203/2006, 2006. p.745-754. ISBN 978-3-540-45764-0.

NOUALI, O.; BLACHE, P. A semantic vector space and features-based approach for automatic information filtering. **Expert Systems with Applications,** v. 26, n. 2, p. 171-179, 2003.

OKAZAKI, N.; ANANIADOU, S. Building an abbreviation dictionary using a term recognition approach. **Bioinformatics,** v. 22, n. 24, p. 3089-3095, 2006. ISSN 1367-4803.

PEREZ-IRATXETA, C.; BORK, P.; ANDRADE, M. A. XplorMed: a tool for exploring MEDLINE abstracts. **Trends in Biochemical Sciences,** v. 26, n. 9, p. 573-575, 2001. ISSN 0968-0004.

PETRIC, I. et al. Literature mining method RaJoLink for uncovering relations between biomedical concepts. **Journal of Biomedical Informatics,** v. 42, n. 2, p. 219-227, 2009. ISSN 1532-0464.

PLAKE, C. et al. ALIBABA: PubMed as a graph. **Bioinformatics,** v. 22, n. 19, p. 2444-2445, October 1, 2006 2006.

POTTENGER, W. M.; YANG, T.-H. Detecting emerging concepts in textual data mining. In: (Ed.). **Computational information retrieval**: Society for Industrial and Applied Mathematics, 2001. p.89-105. ISBN 0-89871-500-8.

PRATT, W.; YETISGEN-YILDIZ, M. **LitLinker: capturing connections across the biomedical literature**. Proceedings of the 2nd international conference on Knowledge capture. Sanibel Island, FL, USA: ACM 2003.

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. 238 (Doutorado). Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, UFSC, Florianópolis (SC).

REBHOLZ-SCHUHMANN, D. et al. EBIMed--text crunching to gather facts for proteins from Medline. **Bioinformatics,** v. 23, n. 2, p. e237-244, January 15, 2007 2007.

RHEINGOLD, H. **La comunidad virtual: una sociedad sin fronteras**. Barcelona: Gedisa Editorial, 1994.

RIJSBERGEN, C. J. V. **Information Retrieval**. Glasgow, Scotland, UK: University of Glasgow, 1979.

ROSE, S. et al. Automatic keyword extraction from individual documents. In: BERRY, M. W. e KOGAN, J. (Ed.). **Text Mining: Applications and Theory**: John Wiley & Sons, Ltd, 2010.

ROSSI, R.; NEVILLE, J. **Modeling the Evolution of Discussion Topics and Communication to Improve Relational Classification**. 1st Workshop on Social Media Analytics. Washington, DC 2010.

RUSSEL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. New Jersey: Prentice-Hall, 1995. 932

SAGA, R. et al. Development and case study of trend analysis software based on FACT-Graph. **Artificial Life and Robotics,** v. 15, n. 2, p. 234-238, 2010. ISSN 1433-5298.

SALTON, G. **Automatic Information Organization and Retrieval**. McGraw Hill Text, 1968. ISBN 0070544859.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing and Management: an International Journal,** v. 24, n. 5, p. 513-523, 1988. ISSN 0306-4573.

SÁNCHEZ, D. A methodology to learn ontological attributes from the Web. **Data & Knowledge Engineering,** v. 69, n. 6, p. 573-597, 2010. ISSN 0169-023X.

SCHREIBER, G. et al. **Knowledge Engineering and Management: The CommonKADS Methodology**. Cambridge, Massachusetts: The MIT Press, 2002.

SHAW, B. **Building a Better Folksonomy: Web-based Aggregation of Metadata**. Technical Report 2005.

ŠILIĆ, A.; DALBELO BAŠIĆ, B. Visualization of Text Streams: A Survey. **Lecture Notes in Computer Science,** v. 6277, p. 31-43, 2010.

SIMON, H. A.; VALDÉS-PÉREZ, R. E.; SLEEMAN, D. H. Scientific Discovery and Simplicity of Method. **Artificial Intelligence,** v. 91, n. 2, p. 177-181, 1997. ISSN 0004-3702.

SMALHEISER, N. R.; TORVIK, V. I.; ZHOU, W. Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. **Computer Methods and Programs in Biomedicine,** v. 94, n. 2, p. 190-197, 2009. ISSN 0169-2607.

SMYTH, B. et al. Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine. **User Modeling and User-Adapted Interaction,** v. 14, n. 5, p. 383-423, 2004. ISSN 0924-1868.

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. ISSN 1532-2882.

STELZL, U. et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. v. 122, n. 6, p. 957-968, 2005. ISSN 0092-8674.

STRÖTGEN, J.; GERTZ, M. TimeTrails: A System for Exploring Spatio-Temporal Information in Documents. **Proceedings of the VLDB Endowment,** v. 3, n. 2, p. 1569-1572, 2010.

STROTGEN, J.; GERTZ, M.; POPOV, P. **Extraction and exploration of spatio-temporal information in documents**. Proceedings of the 6th Workshop on Geographic Information Retrieval. Zurich, Switzerland: ACM 2010.

STUDER, R. et al. Situation and perspective of knowledge engineering. **Knowledge Engineering and Agent Technology: IOS Series on Frontiers in Artificial Intelligence and Applications**, Amsterdam, 2000.

SUBASIC, I.; BERENDT, B. **Web Mining for Understanding Stories through Graph Visualisation**. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining: IEEE Computer Society 2008.

_____. **From bursty patterns to bursty facts: The effectiveness of temporal text mining for news**. 19th European Conference on Artificial Intelligence (ECAI). Lisbon, Portugal: IOS Press. 215**:** 517-522 p. 2010.

SUBAŠIĆ, I.; BERENDT, B. Discovery of interactive graphs for understanding and searching time-indexed corpora. **Knowledge and Information Systems,** v. 23, n. 3, p. 293-319, 2010. ISSN 0219-1377.

SWANSON, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. **Perspectives in Biology and Medicine,** v. 30(1), p. 7-18, 1986.

_____. Medical literature as a potential source of new knowledge. **Journal of the Medical Library Association,** v. 78, n. 1, p. 29-37, 1990. ISSN 0025-7338.

SWANSON, D. R.; SMALHEISER, N. R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence,** v. 91, n. 2, p. 183-203, 1997. ISSN 0004-3702.

TAN, A.-H. **Text Mining: The state of the art and the challenges**. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases 65-70 p. 1999.

TANG, H.; TAN, S.; CHENG, X. A survey on sentiment detection of reviews. **Expert Systems with Applications,** v. 36, n. 7, p. 10760-10773, 2009. ISSN 0957-4174.

TANG, J.; ZHANG, J. Modeling the evolution of associated data. **Data & Knowledge Engineering,** v. 69, n. 9, p. 965-978 2010. ISSN 0169-023X.

TERRA, J. C. C. **Gestão do Conhecimento e E-learning na prática** Rio de Janeiro: Elsevier, 2003.

THEOBALD, M.; SHAH, N.; SHRAGER, J. **Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB**. 2009 AMIA Summit on Translational Bioinformatics. Grand Hyatt, San Francisco 2009.

TSURUOKA, Y.; TSUJII, J.; ANANIADOU, S. FACTA: a text search engine for finding associated biomedical concepts. **Bioinformatics,** v. 24, n. 21, p. 2559-2560, November 1, 2008 2008.

TURNEY, P. D. **Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities**. Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3). LINGUISTICS, A. F. C. Barcelona, Spain**:** 239-24 p. 2004.

TURNEY, P. D.; LITTMAN, M. L. **Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus**. 2002

VAN DER EIJK, C. C. et al. Constructing an associative concept space for literature-based discovery. **Journal of the American Society for Information Science and Technology,** v. 55, n. 5, p. 436-444, 2004. ISSN 1532-2882.

VAN HAAGEN, H. H. H. B. M. et al. Novel Protein-Protein Interactions Inferred from Literature Context. **PLoS ONE,** v. 4, n. 11, p. e7894, 2009.

VECHTOMOVA, O.; ROBERTSON, S.; JONES, S. Query Expansion with Long-Span Collocates. **Information Retrieval,** v. 6, n. 2, p. 251-273, 2003. ISSN 1386-4564.

WANG, X. et al. **Mining common topics from multiple asynchronous text streams**. Proceedings of the Second ACM International Conference on Web Search and Data Mining. Barcelona, Spain: ACM 2009.

WANG, Y.; VECHTOMOVA, O. **Exploring the Use of Term Proximity in Collocate-ranking for Query Expansion**. Joint ACH/ALLC (Association for Computers and the Humanities/Association for Literary and Linguistic Computing) Victoria, BC, Canada 2005.

WEEBER, M. Advances in Literature-based Discovery. **Journal of the American Society for Information Science and Technology,** v. 54, n. 10, p. 913-925, 2003.

WEEBER, M. et al. 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. ISSN 1532-2882.

WEGRZYN-WOLSKA, K. et al. Automated Ontology Learning and Validation Using Hypothesis Testing. In: (Ed.). **Advances in Intelligent Web Mastering**: Springer Berlin / Heidelberg, v.43, 2007. p.130-135. (Advances in Soft Computing).

WEISZ, J.; ROCO, M. C. **Redes de pesquisa e educação em engenharia nas américas**. Rio de Janeiro: FINEP, 1996.

WENGER, E.; MCDERMOTT, R.; SNYDER, W. **Cultivating communities of practice. A guide to managing knowledge**. Harvard Business School Press, 2002.

WITTEN, I. H. et al. **Text Mining: A New Frontier for Lossless Compression**. Proceedings of the Conference on Data Compression: IEEE Computer Society 1999.

WREN, J. D. et al. Knowledge discovery by automated identification and ranking of implicit relationships. **Bioinformatics,** v. 20, n. 3, p. 389-398, 2004. ISSN 1367-4803.

YAN, Y.; MATSUO, Y.; ISHIZUKA, M. **An Integrated Approach for Relation Extraction from Wikipedia Texts**. Online Proc. WWW2009 Workshop on Content Analysis in the WEB2.0 (CAW2.0 2009). Madrid, Spain**:** 7 p. 2009.

YANG, C. C.; SHI, X.; WEI, C.-P. Discovering event evolution graphs from news corpora. **Trans. Sys. Man Cyber. Part A,** v. 39, n. 4, p. 850-863, 2009. ISSN 1083-4427.

YANG, H.; CALLAN, J. **Feature selection for automatic taxonomy induction**. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, MA, USA: ACM 2009.

YONGHUI, W. et al. On-line Hot Topic Recommendation Using Tolerance Rough Set Based Topic Clustering. **Journal of Computers,** v. 5, n. 4, 2010.

YOON, B.; PARK, Y. A text-mining-based patent network: Analytical tool for high-technology trend. **The Journal of High Technology Management Research,** v. 15, n. 1, p. 37-50, 2004. ISSN 1047-8310.

YU, H.; HATZIVASSILOGLOU, V. **Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences**. Proceedings of the 2003 conference on Empirical methods in natural language processing - Volume 10: Association for Computational Linguistics 2003.

ZHANG, K.; ZI, J.; WU, L. G. **New event detection based on indexing-tree and named entity**. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, The Netherlands: ACM 2007.

ZHOU, W.; TORVIK, V. I.; SMALHEISER, N. R. ADAM: another database of abbreviations in MEDLINE. **Bioinformatics,** v. 22, n. 22, p. 2813-2818, 2006. ISSN 1367-4803.

ZHOU, X.; PENG, Y.; LIU, B. Text mining for traditional Chinese medical knowledge discovery: A survey. **Journal of Biomedical Informatics,** v. 43, n. 4, p. 650-660, 2010. ISSN 1532-0464.

ZHU, J. et al. CORDER: COmmunity relation discovery by named entity recognition. K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture, 2005. ACM Press. p.219-220.

_____. Relation discovery from web data for competency management. **Web Intelligence and Agent Systems,** v. 5, n. 4, p. 405-417, 2007. ISSN 1570-1263.

# APÊNDICE A – LISTA DE PUBLICAÇÕES

**Artigos completos publicados em periódicos**

VAN HAAGEN, H. H. H. B. M. ; HOEN, P. B. A. C. ; BOVO, A. B. ; MORREE, A. ; MULLIGEN, E. M. ; CHICHESTER, C. ; KORS, J. A. ; DUNNEN, J. T. ; OMMEN, G. J. B. ; MAAREL, S. M. ; KERN, V. M. ; MONS, B. ; SCHUEMIE, M. J. ; RUTTENBERG, A. . Novel Protein-Protein Interactions Inferred from Literature Context. Plos One, v. 4, p. e7894, 2009.

BALANCIERI, R. ; BOVO, A. B. ; KERN, V. M. ; PACHECO, R. C. S. ; BARCIA, R. M. . A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. Ciência da Informação (Impresso), Brasília, v. 34, n. 1, p. 64-77, 2005.

**Trabalhos completos publicados em anais de congressos**

GONÇALVES, A. L. ; BEPPLER, F. D. ; BOVO, A. B. ; KERN, V. M.; PACHECO, R. C. S. . A Text Mining Approach towards Knowledge Management Applications. In: CRIS-IR, 2006, Copenhagen. Proceedings of the International Workshop on Information Retrieval on Current Research Information Systems, 2006.

GONÇALVES, A. L. ; BEPPLER, F. D. ; GUERIOS, M. C. ; BOVO, A. B. ; IGARASHI, W. ; BORDIN, A. S. ; TCHOLAKIAN, A. B. . Um Modelo Baseado em Mineração de Textos Voltado a Aplicações de Gestão do Conhecimento. In: KM Brasil, 2005, São Paulo. KMBrasil 2005 - O Diálogo Universidade-Empresa na Sociedade do Conhecimento, 2005. v. 11.

**Capítulo de Livro**

KERN, V. M. ; GONÇALVES, A. L. ; BOVO, A. B. . A engenharia do conhecimento e as nuvens de termos aplicadas à análise da pós-graduação interdisciplinar. In: Arlindo Philippi Jr;Antônio J. Silva Neto. (Org.). Interdisciplinaridade em ciência, tecnologia & inovação. Barueri (SP): Manole, 2010.