Marina Carradore Sérgio

UMA ARQUITETURA DE DESCOBERTA DE CONHECIMENTO BASEADA NA CORRELAÇÃO E ASSOCIAÇÃO TEMPORAL DE PADRÕES TEXTUAIS

Trabalho de Conclusão de Curso submetido à Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharel em Tecnologias da Informação e Comunicação.
Orientador: Prof. Dr. Alexandre Leopoldo Gonçalves.

Araranguá
2013

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

Sérgio, Marina Carradore
UMA ARQUITETURA DE DESCOBERTA DE CONHECIMENTO BASEADA NA CORRELAÇÃO E ASSOCIAÇÃO TEMPORAL DE PADRÕES TEXTUAIS / Marina Carradore Sérgio ; orientador, Alexandre Leopoldo Gonçalves - Florianópolis, SC, 2013.
125 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá. Graduação em Tecnologias da Informação e Comunicação.

Inclui referências

1. Tecnologias da Informação e Comunicação. 2. Descoberta de Conhecimento. 3. Bases Textuais. 4. Relacionamentos Indiretos. 5. Computação Distribuída. I. Leopoldo Gonçalves, Alexandre . II. Universidade Federal de Santa Catarina. Graduação em Tecnologias da Informação e Comunicação. III. Título.

Marina Carradore Sérgio

UMA ARQUITETURA DE DESCOBERTA DE CONHECIMENTO BASEADA NA CORRELAÇÃO E ASSOCIAÇÃO TEMPORAL DE PADRÕES TEXTUAIS

Esta Monografia foi julgada adequada para obtenção do Título de "Bacharel em Tecnologias da Informação e Comunicação", e aprovada em sua forma final pelo Curso de Graduação em Tecnologias da Informação e Comunicação.

Araranguá, 12 de Julho de 2013.

_____

Prof. Vilson Gruber, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Alexandre Gonçalves (orientador)

Prof. Juarez B. da Silva

Prof. Anderson Perez

Este trabalho é dedicado a todos que direta ou indiretamente contribuíram em minha formação acadêmica.

AGRADECIMENTOS

Agradeço a todos que contribuíram no decorrer desta jornada, em especial:

À Deus, a quem devo o dom da vida.

À minha família pelo amor e todo o apoio fornecido.

Ao Gustavo pelo companheirismo e compreensão nos momentos de dificuldade.

Ao Orientador Prof. Dr. Alexandre Leopoldo Gonçalves pelo estímulo, dedicação e oportunidades que muito agregaram em minha formação acadêmica.

Aos meus amigos pelo companheirismo e incentivo durante esta jornada e por terem se mantido presentes mesmo com a minha ausência.

Grandes realizações não são feitas por impulso, mas por uma soma de pequenas realizações!

Vicent Van Gogh

RESUMO

Atualmente, o volume de informação gerado aumenta em escala exponencial, sendo que grande parte desta informação encontra-se na forma textual. Através deste formato é possível extrair ativos de conhecimento, ou seja, regras, padrões, tendências, redes, capazes de auxiliar no processo de tomada de decisão dentro das organizações com o intuito de gerar vantagem competitiva. Em virtude da grande disponibilidade de documentos textuais, seja na web ou mesmo nas organizações, assim como, a falta de padronização dos mesmos, tal tarefa constitui-se em um desafio computacional. Neste sentido, é necessário o devido pré-processamento e adequação dos dados. Um meio de se extrair tais ativos de conhecimento é através do processo de Descoberta de Conhecimento em Textos. A partir disto, propõem-se neste trabalho uma arquitetura para descoberta de conhecimento em bases textuais que seja capaz de revelar relacionamentos diretos e indiretos entre padrões textuais (termos) e que tenha suporte da Computação Distribuída. A demonstração de viabilidade é realizada através de um protótipo desenvolvido com base na arquitetura proposta. Como principal resultado do trabalho menciona-se a apresentação da interconexão temporal entre termos através do conceito de associação indireta e posteriormente correlação (associação direta). Além disto, pode-se afirmar que, tanto as distribuições de frequência de um termo quanto os mapas de tópicos, ambos baseados na dimensão tempo, auxiliam no entendimento de determinado domínio do problema. Por fim, a aplicação do protótipo em um cenário permitiu demonstrar que a arquitetura proposta neste trabalho é capaz de atingir resultados consistentes e satisfatórios no que se refere ao entendimento de determinado domínio a partir bases textuais.

Palavras-chave: Descoberta de Conhecimento; Bases Textuais; Relacionamentos Indiretos; Computação Distribuída.

ABSTRACT

Currently the volume of information generated increases in exponential scale. Much of this information is in natural language. Through this format is possible to extract knowledge able to assist the decision making process within organizations in order to generate competitive advantage. Due to the wide availability of textual documents on the web or even in organizations and the lack of standards about document structures such task is a computational challenge. Thus, it is required a suitable data pre-processing. A way to extract such knowledge assets is through the Knowledge Discovery in Texts process. Take it into account we propose in this work an architecture supported by distributed computing for knowledge discovery in textual databases which be able to reveal direct and indirect relationships between textual patterns (terms). The demonstration of feasibility is carried out by a prototype based on the proposed architecture. The main result of this work refers to the demonstration of temporal interconnections among terms through the concepts of indirect association and subsequently correlation (direct association). Moreover, it can be stated that the frequency distributions of a term and topic maps, both based on the temporal vision, help in the understanding of a specific domain problem. Finally, the prototype applied in a scenario has demonstrated that the proposed architecture is able to achieve consistent and satisfactory results towards the understanding of a given domain.

Keywords: Knowledge Discovery; Textual Bases; Indirect Relationships; Distributed Computing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

API – Application Programming Interface.
CPU – Central Processing Unit.
DBL – Descoberta Baseada em Literatura.
DW – Data Warehouse.
EC – Engenharia do Conhecimento.
HTTP – Hypertext Transfer Protocol.
IP – Internet Protocol.
JDBC – Java Database Connectivity.
JSON–JavaScript Object Notation.
KDD – Knowledge Discovery in Databases.
KDT – Knowledge Discovery in Texts.
MD – Mineração de Dados.
MT – Mineração de Textos.
PLN – Processamento em Linguagem Natural.
RAM - Random Access Memory.
UCP – Unidade Central de Processamento.
UFSC – Universidade Federal de Santa Catarina.
XML – eXtensibleMarkupLanguage.

SUMÁRIO

1 INTRODUÇÃO

As evoluções decorrentes das tecnologias da informação e comunicação estão produzindo um avanço significativo no número de documentos que são armazenados e processados. O avanço é decorrente principalmente do aumento da capacidade de processamento, conectividade e armazenamento. Este fenômeno torna-se cada vez mais evidente e vem sendo observado por diversos estudiosos da área.

Uma expressiva quantidade de informações é produzida a todo momento, através de documentos formais ou até mesmo em sites como: Hotmail™, Gmail™, redes sociais, ou de uma maneira mais geral na internet ou na intranet. Conforme Hilbert e López (2011), até o ano de 2007 fomos capazes de gerar aproximadamente 295 exabytes de informação o que equivale a 1024 x 10^6. Para chegar a estes dados os pesquisadores analisaram 60 tecnologias analógicas e digitais de armazenamento, comunicação e computação de informação, durante o período de 1986 a 2007, sendo que dos dados gerados, a humanidade produziu 2,9 x $10^{2º}$ de dados comprimidos, comunicava-se 2 x 10^{21} e realizava 6,4 x 10^{18} instruções por segundo em computadores de uso geral, como por exemplo desktops ou notebooks, sendo que aproximadamente 94% desta informação está contida em meios digitais.

Recentemente o mundo produz entre um e dois exabytes de informação nova por ano, ou seja, algo em torno de 250 megabytes para cada habitante na Terra (LYMAN; VARIAN, 2003). Segundo Bohn e Short (2009), os bytes de informações consumidas por indivíduo nos Estados Unidos têm crescido cerca de 5,4% ao ano, desde 1980, sendo consumido cerca de 3600 exabytes de informações em casas americanas.

Em estudos mais recentes, Bohn e Short (2009), concluíram que as horas dispendidas com o consumo de informação pelos americanos cresceram 2,6% ao ano, no período entre 1980 e 2008, devido ao crescimento populacional e ao tempo dispendido por cada pessoa à busca por informação. Concluíram também que com o advento dos computadores, um terço da informação e mais da metade dos bytes são recebidos de forma interativa.

Gantz e Reinsel (2010), afirmam que até 2020, a quantidade de informações digitais criadas e replicadas no mundo vai crescer aproximadamente 35 trilhões de gigabytes. Os estudos mostraram que em 2009 o universo digital cresceu em torno de 62%, cerca de 800000 petabytes, através desta expressiva quantidade poderíamos fazer uma pilha de DVDs da terra até a lua, ida e volta, afirmam os pesquisadores. No ano de 2010, esse universo alcançaria 1,2 milhões de petabytes, o

que equivale a 10^{15}. Este crescimento explosivo significa que, em 2020, o universo digital será 44 vezes maior, sendo que a pilha de DVDs chegaria agora a meio caminho de marte. Gantz e Reinsel (2010), ressaltam que embora a quantidade de informações no universo digital vai crescer por um fator de 44, e o número de recipientes ou arquivos vai crescer por um fator de 67, no período entre 2009 a 2020, o número de profissionais de TI no mundo vai crescer apenas por um fator de 1,4.

É perceptível o crescente aumento do universo digital. Os documentos impressos, que eram o maior meio de informação textual há algumas décadas, hoje representam apenas 0,003% da informação gerada anualmente (LYMAN; VARIAN, 2003). Segundo Gantz e Reinsel (2010), mais de 70% do universo digital no ano de 2010 foi gerado por usuários que estavam em casa, no trabalho ou até mesmo em movimento, representando cerca de 880 bilhões de gigabytes. Sistemas como o do Facebook™ chegaram a armazenar 1 petabyte de informações no ano de 2008, registrando cerca de 10 bilhões de fotos em seus servidores (WHITE, 2009).

O suporte ao aumento da informação é possível graças à evolução dos meios de armazenamento magnéticos. Segundo Hilbert e López (2011), em 2000 os meios de armazenamento magnéticos representavam 5% da capacidade mundial, saltando para 45% em 2007, e a capacidade de armazenamento per capita que era de 2.866 megabytes em 1993, passou a ser de 44.716 megabytes em 2007.

Com tanta informação disponível é necessário distinguir o que é dado, o que é informação e a partir de que momento a informação torna-se conhecimento. Os computadores trocam dados que podem ser facilmente capturados, comunicados e armazenados. Checkland e Holwell (1998), afirmam que os dados estão associados ao ponto de partida do processo mental, pois é ele que representa um fato existente no mundo real, independentemente de sua relevância ou interesse, visto não haver até aqui relação de fato com o usuário.

Já informação está associada a dados dotados de relevância e propósito. Os seres humanos detêm poder de conferir relevância e propósito, exigindo assim a participação e a análise humana, que pode caracterizar-se como um ponto de vista ou uma forma de representação dos dados em uso (DAVENPORT, 1998; MCGEE e PRUSAK, 1994). McGee e Prusak (1994), comparam o conceito de informação ao conceito de beleza: ambos estão nos olhos do observador. A transformação de dados em informação requer que a pessoa responsável pelo processo decisório receba-os de tal forma que possa relacioná-los e

que atue sobre eles. Assim, a informação deve ser discutida no contexto de usuários e responsáveis por decisões específicas.

A estrutura de apresentação da informação que será armazenada poderá ocorrer de diferentes formas. Dentre estas estruturas estão à estruturada, a semiestruturada e a não estruturada. Em fontes de dados estruturadas de acordo com Silva (2009), os dados estão normalmente dispostos em tabelas, controladas por software de gerenciamento de banco de dados. As informações em fontes semiestruturadas encontram-se entre marcadores (tag), que podem ser reconhecidos ou processados por máquinas, como exemplos estão às páginas HTML e os documentos XML. Já a estrutura de apresentação dos dados obtida através de textos em linguagem natural é conhecida como não estruturada, ou seja, não segue um padrão. Segundo Silva (2009), a informação não estruturada não contém estrutura tabular e nem marcação. Essa informação advém de textos, por exemplo, que o próprio usuário escreve, como um relatório, ou até mesmo um comentário, um post em um blog, enfatizando que em torno de 80% das informações de uma organização estão contidas em documentos textuais, que são a forma mais natural de armazenamento de informações TAN (1999).

A estrutura informacional é um processo de comunicação que visa o conhecimento (BARRETO, 1996; BRAGA, 1996; ZORRINHO, 1995). A distinção entre informação e conhecimento pode se dar no âmbito temporal considerando-se que enquanto a informação é descritiva, isto é, relaciona-se ao passado e ao presente, o conhecimento provê as bases para a previsão do futuro com certo grau de certeza baseado na informação sobre o passado e o presente (KOCK; MCQUEEN; SCOTT, 1997). Ainda na visão de Davenport (1998), o conhecimento é percebido como uma síntese de múltiplas fontes de informação, gerado na mente humana a partir da interpretação ou contextualização de informações externas conjugadas com a própria sabedoria da pessoa.

A rapidez e a magnitude com que o conhecimento gerado passou a ser compartilhado provocou o surgimento de técnicas de reaproveitamento e produção de novos conhecimentos, bem como o aparecimento de novas necessidades de tratar a informação. Entre as possibilidades de ferramental, visando identificar conhecimento, a partir das informações geradas em determinado domínio encontram-se os processos de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database - KDD*) e de Descoberta de Conhecimento em Texto (*Knowledge Discovery in Texts - KDT*).

O processo de KDD é o termo utilizado para promover a descoberta de conhecimento em bases de dados, visando identificar e descobrir relacionamentos implícitos entre as informações armazenadas nos bancos de dados em sistemas organizacionais (SILVA; ROVER, 2011). Quanto ao processo de KDT este é similar ao KDD, porém trabalha com uma coleção de documentos em linguagem natural, buscando padrões e tendências, classificando e comparando documentos (SILVA; ROVER, 2011). Apesar do objetivo em comum, a descoberta de conhecimento, o KDT e o KDD, possuem diferenças importantes. A principal delas refere-se ao tipo de informação, uma vez que KDT trabalha com informações textuais (não estruturadas ou semiestruturadas), enquanto que o KDD trabalha com informações estruturadas, geralmente obtidas a partir de bancos de dados relacionais e/ou orientado a objetos.

Passa-se então a analisar a descoberta de conhecimento sob a ótica de um processo que viabiliza a estruturação da informação vinda de textos. Na visão de Trybula (1999), a descoberta do conhecimento é o processo de transformação de dados em relações previamente desconhecidas e insuspeitas, que podem ser empregadas como previsores de futuras ações. Em virtude dos grandes desafios é necessário encontrar tecnologias capazes de recuperar a informação não estruturada e relevante dos documentos, seguindo critérios definidos pelo usuário e assim obter o melhor retorno, a fim de extrair conhecimento suficiente e tirar vantagem competitiva disto.

A partir de um contexto no qual essas informações estão inseridas estas podem vir a esconder padrões que se relacionam indiretamente e evoluem ao longo do tempo. Os relacionamentos indiretos ocorrem quando os termos apesar de não coocorrerem no mesmo documento possuem elementos textuais em comum. Uma forma de extrair conhecimento em cima dos relacionamentos indiretos é aplicar a técnica de mapa de conceitos e mapas de tópicos sendo que estas são subáreas dos mapas de conhecimento. Segundo Pacheco et al (2007), recentemente com o crescente volume de informações e com o surgimento da área de gestão do conhecimento, as organizações têm procurado aplicar técnicas de mapas de conhecimento para a gestão do conhecimento. Como citado por Eppler (2001), um mapa de conhecimento fornece a orientação para alcançar um determinado universo, ao ajudar a localizar direções, a avaliar situações ou a planejar recursos, sugerindo analogia com um mapa geográfico afirmando que um mapa de conhecimento responde às mesmas quatro perguntas básicas a que um mapa geográfico procura responder: 1. Qual a minha

localização atual? 2. Para onde posso ir? 3. Qual o menor caminho para alcançar o objetivo? 4. Quais os recursos necessários para chegar até o local desejado? A aplicação de gráficos também é utilizada para acompanhar a evolução do relacionamento indireto entre entidades.

Os processos de descoberta de conhecimento com base no contexto em que as informações estão inseridas e nos relacionamentos indiretos que estas podem vir a conter, não são triviais, pois exigem grande demanda computacional em virtude dos algoritmos complexos e análises em cima de uma quantidade extensa de documentos, que contêm vasta informação, seja ela estruturada ou não.

Com isso, este fato constitui-se em desafio, uma vez que tais processos, quando executados a partir de uma infraestrutura computacional inadequada, podem inviabilizar a obtenção de resultados desejados e esperados. Neste cenário, a computação distribuída desempenha um papel fundamental sendo capaz de trabalhar com grandes volumes de dados. Segundo Tanenbaum e Steen (2007), com a computação distribuída é possível utilizar um conjunto de computadores independentes, que na visão do usuário comportam-se como um sistema único e coerente. A principal motivação para a utilização deste sistema é a possibilidade de compartilhar recursos, tais como: componentes de hardware, discos, arquivos e bancos de dados (COULOURIS; DOLLIMORE; KINDBERG, 2005).

Desse modo, o emprego da computação distribuída no processo de extração do conhecimento sobre grandes bases de dados é uma solução plausível que tende a gerar resultados positivos, possibilitando o desenvolvimento de sistemas capazes de analisar grandes fontes de informação com o objetivo de extrair conhecimento que pode a vir a ser utilizado no processo de tomada de decisão.

1.1 PROBLEMÁTICA

A facilidade no acesso aos meios de comunicação, a disseminação da internet e os avanços tecnológicos e computacionais contribuíram de forma significativa para o crescente volume da informação. Com citado acima, mais de 80% das informações de uma organização encontram-se em um formato de dados não estruturado, seja este na forma de relatórios, memorandos, e-mails, documentos.

Utilizando o processo de Descoberta de Conhecimento em Texto (KDT) é possível gerar conhecimento a partir de bases textuais. Contudo, este processo é custoso e não trivial devido principalmente ao

formato como a informação é apresentada, ou seja, não possui estrutura e está sujeita a ambiguidade. Devido ao volume de informações surge também a necessidade de dispor de uma infraestrutura capaz de lidar com o problema de maneira adequada.

Além disto, as informações textuais possuem um caráter temporal sendo necessário lidar com esta característica caso se deseje descobrir comportamentos que descrevam fatos já ocorridos ou que possam vir a ocorrer. Neste sentido, além da infraestrutura computacional adequada deve-se pensar em estruturas capazes de armazenar as informações que passam a ter um contexto multidimensional.

Desse modo tem-se como objetivo de pesquisa: "Como propor uma arquitetura computacional utilizando como base as técnicas de correlação e associação entre elementos textuais e a computação distribuída para a descoberta de relacionamentos indiretos entre padrões textuais considerando a temporalidade?".

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Propor uma arquitetura computacional com base na computação distribuída e em técnicas de associação de elementos textuais que possibilite a descoberta de relacionamentos indiretos e temporais entre padrões textuais.

1.2.2 Objetivos Específicos

Visando atingir o objetivo principal, alguns objetivos específicos são requeridos, entre eles:

- Analisar o panorama atual dos processos de descoberta de conhecimento;
- Estudar modelos que suportem a representação de informações que relacionem padrões de maneira temporal;
- Desenvolver um protótipo voltado à identificação de relacionamentos indiretos considerando a temporalidade que justifique a proposição de arquitetura;

- Realizar uma discussão dos resultados oriundos do processo de descoberta de conhecimento.

1.3 METODOLOGIA

O trabalho será desenvolvido com base um uma pesquisa exploratória, através do desenvolvimento de um protótipo de descoberta de conhecimento, a partir de bases textuais. A metodologia de desenvolvimento deste trabalho é dividida em cinco etapas:

Etapa 1: análise da literatura focando nas seguintes áreas: descoberta de conhecimento e computação distribuída.

Etapa 2: proposição da arquitetura lógica, modelagem da camada de persistência e desenvolvimento da arquitetura física com suporte para o processamento distribuído.

Etapa 3: implementação das camadas de persistência e de aplicação de descoberta de conhecimento, de modo que utilize a arquitetura física construída anteriormente.

Etapa 4: desenvolvimento e teste de um protótipo considerando um cenário de uso.

Etapa 5: avaliação dos resultados obtidos através da utilização da arquitetura proposta neste trabalho.

1.4 ORGANIZAÇÃO DO TEXTO

O documento está dividido em seis capítulos. No presente capítulo apresenta-se o projeto, expondo uma breve contextualização e apresentando a problemática vislumbrada, assim como os objetivos gerais e específicos.

No segundo capítulo é realizada uma revisão sobre a área de descoberta de conhecimento promovendo um maior detalhamento do processo da Descoberta de Conhecimento em Texto com base em formatos não estruturados.

O terceiro capítulo realiza uma breve revisão sobre a área de Computação Distribuída visto que esta fornece a base para a arquitetura proposta e o desenvolvimento do protótipo.

O quarto capítulo propõe uma arquitetura de descoberta de conhecimento, a partir de bases textuais aliada à computação distribuída,

para a descoberta de relacionamentos indiretos e temporais. Este capítulo divide-se em duas partes, sendo: (a) um detalhamento do modelo lógico da arquitetura; e (b) uma descrição dos componentes tecnológicos e serviços da arquitetura (modelo físico).

O quinto capítulo apresenta e discute os resultados obtidos assim como as possibilidades de análise considerando a proposição do trabalho.

Por fim, o sexto capítulo contém as considerações finais e os trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO

Neste capítulo será realizada uma explanação sobre a área de pesquisa do trabalho. Contextualizar-se-á a evolução histórica envolvendo a pesquisa e os principais termos contidos na mesma.

Com o advento da digitalização dos documentos e o desenvolvimento das redes de computadores, o volume de informação aumentou, chegando ao ponto em que a escala de apreensão humana não conseguiu mais absorver, gerando assim um lapso entre a criação dos dados e a compreensão dos mesmos (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992). O crescente número na quantidade de informação que vem sendo gerada chega à escala exponencial (GREENGRASS, 2000; KOBAYASHI; TAKEDA, 2000; LYMAN, 2000; HIMMA, 2007).

Quando devidamente tratada a informação passa a ser um instrumento de auxílio importante para a tomada de decisão dentro das organizações. Só é extraída informação dos documentos, caso os mesmos estejam com os dados devidamente estruturados. Dados de acordo com Tuomi (1999), são fatos simples que podem vir a se transformar em informação caso contenham uma estrutura de apresentação. A informação passa a se tornar conhecimento quando esta é interpretada, inserida em um contexto, ou quando é adicionado significado a mesma.

Os dados acabam por se tornar um pré-requisito para a informação, e a informação é imprescindível para a obtenção do conhecimento, sendo que para a mesma conseguir alcançar o estágio de conhecimento esta passa por etapas envolvendo meios computacionais, onde ocorre o processamento e são aplicadas técnicas de mineração de dados. Davenport (1998), apresenta a informação como uma ponte entre dados brutos e conhecimento que eventualmente possa se obter. A Figura 1 representa a relação entre dado, informação e conhecimento fazendo um paralelo com o valor agregado que envolve cada nível em que a informação encontra-se.

Figura 1 – Dado X Informação X Conhecimento.

Fonte: Adaptação da Pirâmide de Maslow [1].

Descobrir conhecimento representa o processo de identificar, receber informações relevantes e através disto poder computá-las e agregá-las ao nosso conhecimento prévio, mudando o estado do conhecimento atual, a fim de que determinada situação ou problema possa ser solucionado (WIVES, 2004). A gestão correta do conhecimento, torna-se uma grande aliada para as organizações que buscam obter vantagens trabalhando em cima da informação. De acordo com Fialho et al. (2006), a gestão do conhecimento trata da prática de agregar valor à informação e difundi-la, tendo como objeto central o aproveitamento dos recursos existentes na empresa. Ainda segundo O'Leary (1998) e Steels (1993), o gerenciamento do conhecimento permite promover o crescimento dos negócios, a comunicação e a preservação do conhecimento no meio organizacional.

Para Moresi (2000), a gestão efetiva de uma organização requer a percepção objetiva e precisa dos valores da informação e do sistema de informação. Em decorrência disto muitos são os desafios enfrentados pela área responsável pela extração de informação, que trabalha com um conjunto de técnicas que têm como objetivo extrair de fontes

[1] Teoria da motivação humana proposta por Abraham Maslow, segundo a qual as necessidades humanas estão organizadas e dispostas em níveis, numa hierarquia de importância e de influências. Na base da pirâmide situam-se as necessidades fisiológicas, sobre estas as necessidades de segurança, seguidas pelas necessidades sociais, necessidades de estima e chegando ao topo, as necessidades de auto realização (CHIAVENATO, 2003).

semiestruturadas ou não estruturadas, informação selecionada (ETZIONI, 2008).

A maior parte da informação encontra-se em formato não estruturado, ou seja, em linguagem natural. Muitas pesquisas na área da descoberta do conhecimento vêm sendo feitas, a fim de extrair conhecimento em cima das estruturas organizacionais dos dados não estruturados, pois com elas surgem os desafios na hora de coletar, organizar e extrair padrões relevantes que possibilitem auxiliar na tomada de decisão.

As áreas relativas à extração e recuperação da informação e descoberta de conhecimento auxiliam no processo de desenvolvimento dos sistemas na descoberta de conhecimento. 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. Nestas grandes bases textuais estão contidas informações adormecidas, camufladas, até que o minerador as encontre e as transformem em informações preciosas para a organização (RAMOS; BRÄSCHER, 2009). Segundo Bovo (2011), a análise de dados passa a ter um caráter mais exploratório, visando identificar ou explicitar conhecimento oculto nessas bases de dados.

O conhecimento extraído em cima da informação propicia muitas oportunidades para quem deseja extrair algo a mais e utilizar como parâmetro para a tomada de decisão. Por outro lado, de acordo com Levy (2005), o problema de se lidar com muita informação é que se perde um tempo que poderia ser melhor empregado: pensando, contemplando e raciocinando.

2.1 ESTRUTURA DE APRESENTAÇÃO DA INFORMAÇÃO

No âmbito das estruturas como a informação disponível é apresentada trabalha-se com três tipos, sendo elas: a estruturada, a semiestruturada e a não estruturada. A informação estruturada é normalmente disposta em uma tabela, esta tabela por sua vez é gerenciada por um software de controle de banco de dados. As informações em fontes semiestruturadas são dispostas entre marcadores (tag), que podem ser reconhecidos ou processados por máquinas, como exemplos tem-se as páginas HTML e os documentos XML. Esta por sua vez é intermediária entre a informação textual e a informação estruturada, tipicamente encontrada em sistemas de banco de dados relacionais. Já as informações não estruturadas são todos os documentos

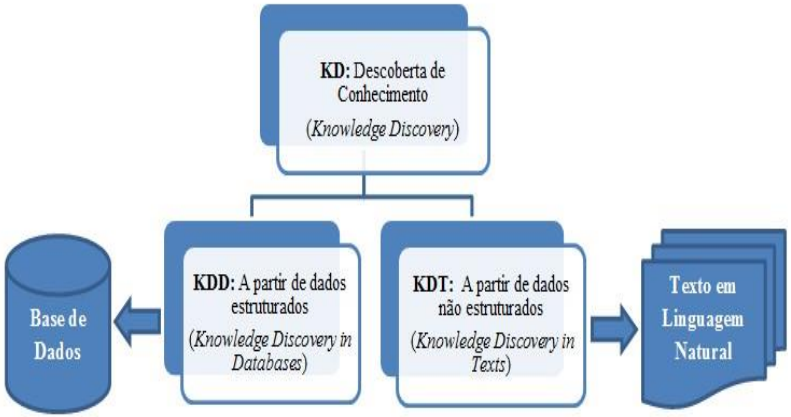
dispostos em linguagem natural, sendo que estes por sua vez, não utilizam nenhum padrão e não seguem uma estrutura na hora de ser elaborado.

A maior parte da informação eletrônica está disponível em bases de dados conhecidas como bases não estruturadas, cujo formato destas, está adequado aos seres humanos que, através da leitura, são capazes de decodificar os dados contidos no texto e compreender a informação ali contida (SCHIESSL, 2007). Para Bovo (2011), a informação não estruturada possui apenas uma estrutura sintática, isso no âmbito da escrita, porém no âmbito da ciência da computação, ela é considerada como não estruturada.

Os sistemas baseados em informação estruturada permitem recuperar informações com maior facilidade, pois os usuários geralmente conhecem a estrutura como os dados são dispostos e há linguagens de consulta disponíveis, obtendo assim resultados mais precisos em cima das consultas elaboradas (JUNQUEIRA, 2009). Caso essa recuperação seja realizada em uma base não estruturada, a dificuldade será maior, devido ao fato de o usuário não conhecer a estrutura dos dados. Por este motivo, a consulta será formulada com base em palavras-chaves, o quê em geral, não produzem resultados tão expressivos. Já em sistemas que trabalham com a recuperação semiestruturada, o usuário geralmente desconhece a estrutura dos dados, e formula consultas que mesclam bases textuais e mecanismos de recuperação estruturada.

A descoberta de conhecimento a partir destas estruturas de apresentação da informação pode ser decomposta em duas vertentes: Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge Discovery in Databases* – KDD), e Descoberta de Conhecimento em Textos (do inglês *Knowledge Discovery in Texts* – KDT). Esta divisão tem como base o conteúdo que será analisado, em que, caso este seja previamente organizado e estruturado o processo de descoberta a ser utilizado será o KDD. Caso o conteúdo encontrar-se disperso em documentos textuais o processo utilizado será o KDT (RAMOS, BRASCHER; 2009). Ambos trabalham com a descoberta de conhecimento, porém a uma importante diferença entre os dois, cujo relacionamento está diretamente ligado ao tipo de informação, uma vez que o KDT trabalha com informações textuais (não estruturadas ou semiestruturadas), enquanto que o KDD trabalha com informações estruturadas, geralmente obtidas a partir de bancos de dados relacionais e/ou orientado a objetos. A Figura 2 explicita está diferença.

Figura 2 – Diferença entre KDD e KDT.



Fonte: Autor.

2.2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

À medida que os avanços tecnológicos vão acontecendo, esforços vão sendo dispendidos para viabilizar a utilização eficiente de grandes volumes de dados e da obtenção da informação. O KDD apresenta-se como uma opção para atender essa necessidade. No processo de KDD os dados estão sempre bem estruturados e dispostos em formatos numéricos ou símbolos preparados para a leitura por computador (WEISS et al., 2005). Para Berry e Linoff (1997), o processo de KDD é a análise e a exploração automática ou semiautomática de grandes quantidades de dados, com a intenção de encontrar regras e padrões significativos.

Para Fayyad et al. (1997), em geral, o campo de pesquisa que envolve o KDD compreende o desenvolvimento de técnicas e métodos que procuram prover significado aos dados. Os métodos clássicos que transformam os dados em informação trabalham na análise manual e na interpretação, porém, vão de encontro com a grande disponibilidade existente nas bases de dados, o que o torna lento, caro e altamente subjetivo o processo. Segundo Schiessl (2007), o processo básico do KDD envolve a tradução da informação em seu nível mais elementar, o dado, geralmente armazenado em grandes bases, em forma mais compactas, mais resumidas e mais úteis. Assim, o KDD é uma tentativa

de lidar com um problema que, na era da informação digital, tornou-se real e visível: a sobrecarga da informação.

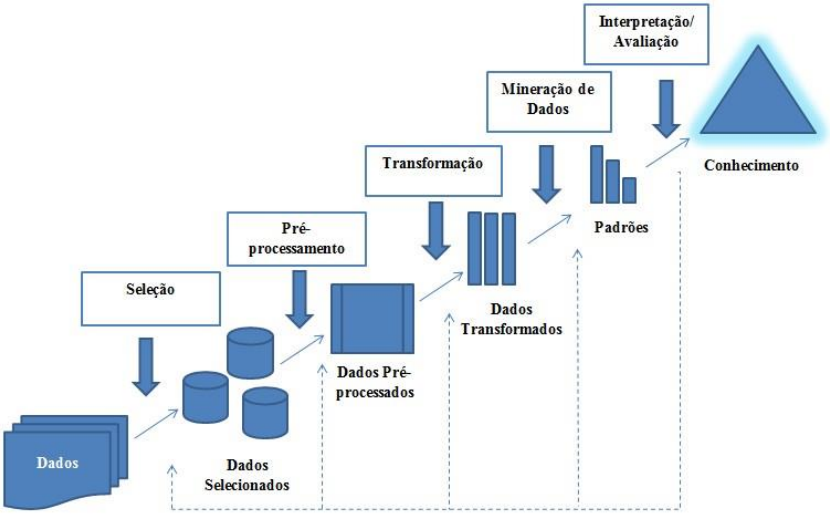
O processo de descoberta de conhecimento em bases de dados compreende a seleção dos dados, o pré-processamento que adequa os dados aos algoritmos, a efetiva mineração dos dados que envolvem o uso de técnicas de mineração, a validação dos resultados e a análise e interpretação dos resultados para aquisição do conhecimento. Em suma o principal objetivo do KDD é a tradução de dados brutos em informações relevantes (VIANNA et al, 2010).

Baseado nos autores (FAYYAD et al, 1997; BERRY e LINOFF, 1997; TRYBULA, 1997; HAN e KAMBER, 2000; KANTARDZIC, 2003; SCHIESSL, 2007), há variações quanto a ordem como as etapas do processo de descoberta de conhecimento em bases de dados são apresentadas, porém será apresentado uma visão geral destas etapas:

- Definição do problema: definição dos objetivos, obstáculos, dificuldades, assim como a identificação das áreas ou pessoas que serão beneficiadas;
- Seleção dos dados: etapa em que os dados da base de dados serão adequados ao processo de coleta;
- Pré-processamento: nesta etapa serão removidos erros ou inadequações dos dados frente ao processo de KDD.
- Análise exploratória: exame dos dados buscando identificar estruturas que expressem alguma relação entre registros ou variáveis;
- Redução de variáveis: redução da dimensionalidade em função da lista extensa de variáveis;
- Relacionamento de objetivos: nesta etapa ocorre a seleção do algoritmo de mineração de dados apropriado às necessidades do problema;
- Mineração de dados: envolve o uso da ferramenta de mineração de dados com o propósito de extrair padrões frente aos dados;
- Interpretação dos resultados: análise dos padrões para verificação de sua utilidade e realimentação da informação;
- Transformação do conhecimento adquirido em ação: utilização do conhecimento extraído ou sua incorporação à base de conhecimentos acumulados.

Na visão de Fayyad et al. (1997) os processos que envolvem o KDD podem ser representados de acordo com a Figura 3.

Figura 3 – Passo a passo das etapas do KDD.



Fonte: Adaptado de (FAYYAD et al, 1997).

2.3 DESCOBERTA DE CONHECIMENTO EM TEXTO

Com a popularização dos meios de comunicação e o acesso facilitado a tecnologia, gerou-se uma explosão de documentos textuais de toda ordem, seja em correspondências eletrônicas, em publicações científicas, ou em sites na internet e intranet, trazendo consigo diversas finalidades. A cada ano são produzidos aproximadamente 968 mil livros, 80 mil revistas, 40 mil periódicos e bilhões de documentos (LYMAN; VARIAN, 2003).

De acordo com Chaves (2007), uma grande parte do conhecimento existente atualmente está na forma de texto (a maioria não estruturado), e por este motivo, esse conhecimento precisa ser identificado, representado e manipulado, de modo a tornar-se realmente útil para as organizações.

Parte considerável desta informação encontra-se na forma de textos nos mais diversos formatos. Desde a década de noventa estudos como os de Wilks e Catizone (1999) e Tan (1999), já apontavam que 80% da informação encontrava-se na forma textual. Com vasta informação disponível surgiu à área voltada à descoberta de conhecimento em texto, com o intuito de extrair informações relevantes em cima destes documentos, que possam vir a se tornar conhecimento

com o auxílio de técnicas, para com isso utilizá-lo ou incorporá-lo a base de conhecimento da organização.

Para Schiessl (2007), a descoberta de conhecimento baseada em texto propõe soluções para tratar a informação eletrônica textual com o auxílio de máquinas, buscando diminuir o impacto da sobrecarga de informação. Na visão de Hearst (1999), os documentos textuais expressão informações preciosas, porém estão codificadas de maneira difícil de ser decifrar automaticamente. Muitas pesquisas vêm sendo feitas a fim de criar técnicas para extrair o conhecimento contido nestas bases.

O KDT possui uma estrutura implícita que necessita de técnicas especializadas para ser reconhecida por sistemas automatizados. O processamento em linguagem natural (PLN) trabalha com estas estruturas implícitas, sendo uma delas a estrutura sintática (RAJMAN; BESANÇON, 1997). Em parceria, a técnica de PLN e o KDD são capazes de transformar os dados textuais em informação para assim possibilitar a aquisição do conhecimento (SCHIESSL, 2007).

Segundo Trybula (1999), o KDT assemelha-se muito ao KDD, porém o KDT foca em documentos textuais. Na visão de Tan (1999) e Feldman et al.(2001), o KDT é a área do KDD que trata dos documentos textuais, sendo que ambos referem-se ao processo de extração de padrões não triviais e de conhecimento útil para determinado objetivo com base em determinados documentos. Entretanto, a área que envolve o KDT torna-se mais complexa devido ao fato de os dados não estarem estruturados e sim em linguagem natural. Para Barion e Lago (2008) o KDT baseia-se em técnicas específicas voltadas ao tratamento de textos, para assim obter conhecimentos implícitos em banco de dados textuais.

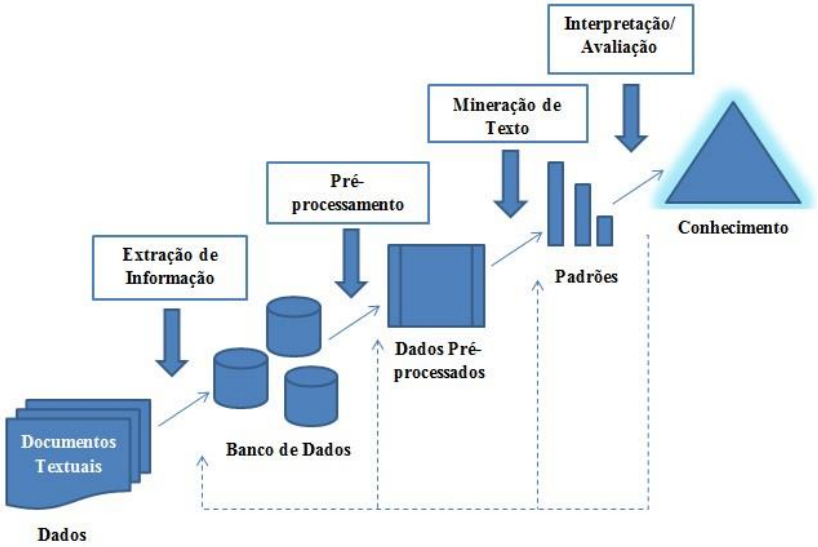
Os passos do KDT possuem adaptações para que possa ser aplicado em informações não estruturadas. Segundo Ceci et al. (2010), as principais diferenças ocorrem nos seguintes passos:

- Extração de Informação: baseados no domínio do problema são selecionados os textos que irão envolver o processo de descoberta de conhecimento em texto.
- Pré-processamento: "o objetivo desta fase é a eliminação de termos não relevantes (*stop-words*), redução das palavras aos seus radicais (*stemming*), correções ortográficas e outros aspectos morfológicos e também sintáticos que as expressões textuais possuem." Nesse sentido, o Processamento de Linguagem Natural (PLN) é

fundamental nesta fase. O processo visa à estruturação da informação com o objetivo de minimizar o custo da alta dimensionalidade. A partir da estruturação da informação já podem ser aplicadas as técnicas que compreendem o KDD.

A Figura 4 fornece uma visão das etapas envolventes no processo de KDT.

Figura 4 – Passo a passo das etapas do KDT.



Fonte: Adaptado de (MOONEY; NAHN, 2005).

2.4 MODELOS BASEADOS EM COOCORRÊNCIA

O processo de descoberta de conhecimento em bases textuais precisa utilizar métodos que forneçam suporte no momento da agregação da informação textual. Os modelos baseados em coocorrência buscam demonstrar a possibilidade de elementos textuais ocorrerem em combinação um com o outro frente a uma coleção de documentos.

Os processos de recuperação de informação tradicionais, assim como a análise da informação que dão suporte a estes modelos de documentos, em geral, enfrentam problemas relacionados à falta de contexto, dificultando assim o processo de recuperação, devido ao fato de muitos termos que ocorrem ao longo da coleção não possuírem relacionamentos entre si GONÇALVES (2006). Em decorrência deste

fator pode-se vir a não ter o real resultado do processo de recuperação e análise. Gonçalves (2006), também afirma que através da utilização de modelos baseados em coocorrência são obtidos resultados satisfatórios tanto na representação de documentos, quanto no mapeamento de conhecimento implícito em bases textuais.

Através de cálculos oriundos da estatística é possível encontrar um grau de relação entre os elementos textuais desejados, tornando assim os cálculos estatísticos fortes aliados do modelo correlacional. Segundo Stevenson (2001), o objetivo do estudo correlacional é a determinação da força do relacionamento entre duas observações emparelhadas. O termo "correlação" significa literalmente "co-relacionamento", pois indica até que ponto os valores de uma variável estão relacionados com os de outra. Ainda segundo Lira (2004), a partir da correlação é possível encontrar o grau de relacionamento entre duas variáveis.

A função de correlação gera como resultado o grau de correlação ou coeficiente de correlação. O conjunto de coeficientes de correlação, forma uma matriz de correlação TERMOxTERMO, ou seja, uma matriz que contém em cada célula o valor correlacional entre dois termos quaisquer representado pelo símbolo W e entre parênteses os termos correlacionados, onde T representa o termo relacionado. A diagonal principal da matriz não é preenchida, devido ao fato da não necessidade de relacionar o termo com ele mesmo, ou seja, não é necessário saber o valor correlacional entre T1 e T1, por ser o mesmo termo.

Tabela 1- Matriz de correlação entre termos.

Termos	T1	T2	T3	T4	T5	T6
T1		W (T1, T2)	W (T1, T3)	W (T1, T4)	W (T1, T5)	W (T1, T6)
T2	W (T2, T1)		W (T2, T3)	W (T2, T4)	W (T2, T5)	W (T2, T6)
T3	W (T3, T1)	W (T3, T2)		W (T3, T4)	W (T3, T5)	W (T3, T6)
T4	W (T4, T1)	W (T4, T2)	W (T4, T3)		W (T4, T5)	W (T4, T6)
T5	W (T5, T1)	W (T5, T2)	W (T5, T3)	W (T5, T4)		W (T5, T6)
T6	W (T6, T1)	W (T6, T2)	W (T6, T3)	W (T6, T4)	W (T6, T5)	

Fonte: Autor.

Segundo Downie e Heath (1959), existem situações em que o relacionamento entre as duas variáveis não é linear, ou uma delas não é contínua, ou o número de pares das medidas é muito pequeno. Então, para cada uma dessas situações há necessidade de uma medida adequada de associação entre as variáveis. A seguir são apresentados os principais modelos baseados em coocorrência, segundo a visão de Gonçalves (2006), assim como os principais conceitos utilizados por esses modelos.

2.4.1 Frequência

Uma das formas mais simples, porém imprecisa, para estabelecer a relação entre dois termos é a utilização do modelo da frequência. Para Gonçalves (2006), O fato de duas palavras, ou qualquer outro termo textual, aparecerem frequentemente em uma determinada coleção de documentos demonstra a evidência de relacionamento entre elas. Schiessl (2007), afirma que o modelo por frequência considera a quantidade de vezes que o termo aparece no decorrer do documento. Entretanto a utilização deste método em documentos textuais sem um devido tratamento é preocupante, em decorrência de o próprio idioma conter muitos artigos e preposições que ocorrem frequentemente no

documento. Uma das soluções encontradas para a eliminação e tratamento dos artigos e preposições é a utilização de uma lista de controle contendo os termos a serem retirados, conhecido como (*stop list*).

2.4.2 Média e Variância

O modelo média e variância permite encontrar relações de uma maneira mais objetiva, pois considera a relação de palavras mesmo havendo uma distância distinta entre elas no texto, afinal a quantidade de palavras que aparece entre outras duas palavras varia, ou seja, caso deseja-se saber a distância entre dois termos T1 e T2 basta efetuar um processo onde a quantidade de palavras entre os dois termos será computada, processo conhecimento como utilização de janelas que nada mais é do que a quantidade de palavras em cada um dos lados de uma determinada palavra.

O cálculo deste modelo é realizado da seguinte maneira: primeiramente é calculada a média das distâncias em que as palavras ocorrem no texto através da seguinte equação:

$$\bar{\mathrm{d}} = \frac{\mathrm{s}}{\mathrm{n}}$$

Onde s representa a soma das distâncias, e n o número de coocorências dos termos.

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

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

Onde *n* é o número de vezes que as duas palavras coocorrem, *di* é a distância da *ith* coocorrência, e $\bar{\mathrm{d}}$ é a média das distâncias. Caso as distâncias sejam sempre as mesmas, a variância será zero. Caso contrário, se as distâncias acontecem aleatoriamente, ou seja, não configuram um padrão de relacionamento, a variância será alta. Finalmente, é realizado o cálculo de desvio padrão:

$$\mathrm{S} = \sqrt{s^2}$$

Sendo que, valores de desvios altos indicam relacionamentos pouco relevantes. A raiz quadrada da variância é utilizada para avaliar os vários deslocamentos entre duas palavras ou qualquer outra estrutura textual. A média e a variância podem determinar o grau de relacionamento entre os termos.

2.4.3 Teste de hipótese

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, a hipótese nula H0 indica a ausência de associação entre os dois termos, além das ocorrências ao acaso. Para tal, calcula-se a probabilidade p que o evento ocorreria se H0 fosse verdadeira, rejeitando-se H0 se p é muito baixa, normalmente caso seja abaixo de um nível de significância de $p < 0.05$, 0.01, 0.005, ou 0.001), caso contrário, aceita-se H0 como sendo possível (BOVO, 2011). Assim, quando a hipótese nula é rejeitada há probabilidade da existência do relacionamento entre as duas palavras além das ocorrências ao acaso e, de maneira similar, quando se aceita a hipótese nula, acredita-se que não existe um relacionamento entre as duas palavras. Caso a probabilidade p seja superior ao nível de significância, H0 não pode ser rejeitado.

2.4.4 Teste de Pearson – chi-square (x^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 (BOVO, 2011). A aplicação do Teste de Pearson utiliza uma tabela 2*2 (tabela de contingência), conforme observado na Tabela 2.

Tabela 2- Tabela de contingência de 2x2.

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

Fonte: Autor.

Sendo que a representa a quantidade de vezes em que w_1 e w_2 ocorrem conjuntamente, b indica a quantidade de ocorrências de w_1 sem a presença de w_2, c indica a quantidade de ocorrências de w_2 sem a presença de w_1, e d é o tamanho da coleção de documentos menos o número de documentos que não contenham w1 e/ou w2, sendo d = N - a - b - c, onde N é o tamanho da base (GONÇALVES, 2006).

2.4.5 Phi-squared (ϕ^2)

O phi-squared também utiliza uma tabela de contingência, similar ao método anterior. Segundo Conrad e Utt (1994), o Phi-squared tende a favorecer associações com alta frequência. O Phi-squared (CHURCH; GALE, 1991) é definido como:

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}, \text{ onde } 0 \leq \phi^2 \leq 1.$$

2.4.6 Informação Mútua

Segundo Church e Hanks (1990), a Informação Mútua compara a probabilidade de um par de palavras (ou qualquer outra unidade linguística) aparecerem 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 o índice apurado através da Informação Mútua será um valor negativo. Para Schiessl (2007), a informação mútua indica a proximidade da distribuição dos documentos que contém o termo com a distribuição dos documentos que estão contidos na coleção.

A fórmula padronizada para o cálculo de MI é definida como:

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

Onde P(x, y) é a probabilidade de duas palavras x e y ocorrerem conjuntamente, P(x) e P(y) são as probabilidades de x e y ocorram individualmente, e N, o tamanho da base. Quando existe um

relacionamento forte entre as palavras (características), I(x, y) será maior que 0.

2.5 ASSOCIAÇÃO DE ELEMENTOS TEXTUAIS

O processo que envolve a correlação trata do relacionamento direto existente entre os termos contidos em uma coleção de documentos. Por outro lado, a associação mostra o relacionamento indireto entre dois elementos textuais, baseado nos contextos nos quais estão inseridos. Para obter os cálculos referentes a este processo, é necessário utilizar primeiramente os modelos baseados em coocorrência para, numa etapa posterior, buscar identificar relacionamentos indiretos entre elementos que não coocorrem, ou que coocorrem com uma frequência muito baixa, através do contexto de cada um (BOVO, 2011). As áreas da biomedicina e bioinformática vêm provocando grandes avanços envolvendo a associação entre elementos textuais. Os trabalhos relativos a estes métodos estão mais relacionados à área de Descoberta Baseada em Literatura (DBL).

A Descoberta Baseada em Literatura (DBL – *Literature-Based Discovery*) tem por objetivo descobrir relacionamentos implícitos na bibliografia científica, a fim de gerar potenciais hipóteses para novas descobertas (SMALHEISER, 2011). Para Tardelli et al (2002), a DBL envolve formulações de hipóteses científicas através da busca de conexões entre estruturas de conhecimento publicamente disponíveis, mas impensadas, isto é, jamais mencionadas ou aplicadas juntas. Segundo Bovo (2011), a DBL utiliza métodos de mineração de texto para a descoberta de novos conhecimentos através de relacionamentos indiretos entre elementos textuais. A associação de elementos textuais ocorre quando entidades se relacionam de forma indireta.

Os primeiros passos envolvendo a DBL foram dados por um cientista norte americano, Don R. Swanson, ao qual este focou em pesquisas que envolvem a área biomédica e a descoberta de relacionamento implícito entre padrões.

Swanson realizou extensas pesquisas em bases de dados biomédicas em busca de informações sobre a síndrome de Raynaud, que posteriormente veio a resultar na conexão da mesma com o óleo de peixe (SWANSON, 1986). Os pacientes que possuíam a síndrome apresentavam alterações no sangue, como “alta viscosidade” e “elevada agregação de plaquetas”. A partir deste momento, Swanson realizou uma revisão da literatura envolvendo a alta viscosidade do sangue. O

pesquisador descobriu uma conexão entre este termo e o termo “óleo de peixe”, apontando que o óleo de peixe provocava a diminuição da viscosidade do sangue e a agregação de plaquetas. Assim concluiu que havia um relacionamento indireto entre a doença “Síndrome de Raynaud” e os fatores “alta viscosidade do sangue” e “elevada agregação de plaquetas”. Swanson gerou a seguinte hipótese: “óleo de peixe” pode ser útil para reduzir a “alta viscosidade do sangue” e a “elevada agregação de plaquetas” em seres humanos e então amenizar os sintomas da “Síndrome de Raynaud”. Mais tarde, Swanson também encontrou um relacionamento entre os termos “Magnésio” e “Enxaqueca”, onde a deficiência de magnésio exerce influência na enxaqueca (SWANSON, 1988).

Para Barçante (2011), as descobertas de Swanson sobre inferências é uma comprovação de uma hipótese lançada pelo usuário que é testada nas buscas de informação, concluindo que o computador realmente não fez a “descoberta”, mas a comprovou.

A Figura 5 demonstra melhor a diferença entre correlação e associação de elementos textuais. Observam-se dois elementos centrais que são Vitor e Laura, dos quais serão os objetos de estudo. Estes por sua vez poderiam ser palavras, documentos, ou qualquer outro elemento dependendo do contexto ao qual estejam inseridos. Tais elementos são genericamente tratados como entidades. Vitor possui uma rede de relacionamentos com Ana, João e Mel, que são correlacionados a Vitor. Já no contexto que envolve Laura, a mesma relaciona-se com Maria, Alex e Joana. Observar-se ainda que Ana está relacionada a Vitor porém conhece Maria e Maria é amiga de Laura. Neste caso, existe uma correlação entre os termos Ana e Maria. Outra característica, Pedro se relaciona tanto com Vitor quanto Laura, evidenciando algo em comum entre as entidades, o que representa a associação.

Figura 5 – Associação e coocorrência de termos.

Fonte: Autor.

Em virtude da associação e em decorrência da escala temporal, Vitor que antes era independente de Laura e vice versa, podem passar a ser mencionados em conjunto dentro de um determinado contexto.

A descoberta de relacionamentos indiretos entre entidades pode conferir informações importantes e preciosas dependendo do contexto ao qual estão inseridas e como serão tratadas essas informações. Segundo Gonçalves et al. (2005), a análise de relacionamentos indiretos podem revelar padrões mais complexos entre as entidades promovendo diferente perspectivas na análise de relações.

Para analisar o contexto da imagem perante aos meios computacionais serão criados dois vetores de contexto. Para Bovo (2011), os vetores de contexto são utilizados para a obtenção da lista dos conceitos que estão fortemente relacionados a um dado conceito. Esse contexto também pode ser dividido por tempo, nesse caso têm-se vetores temporais de contexto. A partir destes vetores é possível utilizar a DBL e evidenciar associações entre elementos textuais. A Tabela 3 representa os vetores de contexto de Vitor e Laura com seus respectivos objetos correlacionados, respectivamente, em primeiro momento, contendo o valor correlacional existente entre as entidades em questão e as demais.

Tabela 3- Vetores de contexto.

	Ana	João	Mel	Pedro	Maria	Alex	Joana	Vitor	Laura
Vitor	**0,1**	**0,2**	**0,1**	**0,4**	**0**	**0**	**0**	**-**	**0**
Laura	**0**	**0**	**0**	**0,5**	**0,2**	**0,1**	**0,3**	**0**	**-**

Fonte: Autor.

Em um segundo momento o contexto das entidades em questão pode expor um novo cenário, em virtude da escala temporal, como mostra a Tabela 4, correspondente a entidade Vitor.

Tabela 4- Matriz de dados representando o vetor de contexto Vitor representado em escala temporal.

Termo de Origem	Termos Correlacionados	Peso (Tempo 1)	Peso (Tempo 2)
V1	Ana	W1	W1
V1	João	W2	W2
V1	Mel	W3	W3
V1	Pedro	W4	
V1	Alex		W5

Fonte: Autor.

Na primeira coluna da matriz temos a entidade em questão, Vitor. A segunda coluna representa o vetor de contexto. A terceira coluna mostra o peso de correlação entre os termos relacionados, representado pela letra W, em um determinado período. Este período representa uma escala temporal que pode ser medida em horas, semanas, meses, enfim, dependendo do cenário ao qual se pretende avaliar. A quarta coluna remete a uma evolução temporal, ou seja, o peso da correlação entre os elementos em um segundo momento. À medida que o tempo passa

novos elementos podem passar a ser correlacionados com a entidade em questão, no caso Alex, e outros deixarem de serem mencionados, como é o caso de Pedro, conforme remete a Tabela 4.

Para analisar os relacionamentos indiretos observa-se os termos em comum entre as entidades e a evolução temporal entre os termos relacionados. Segundo Pacheco et al (2007), existem maneiras de extrair conhecimento em cima das informações obtidas, uma forma é aplicar a técnica de mapa de conhecimento para a gestão do conhecimento. Como citado por Eppler (2001), um mapa de conhecimento fornece a orientação para alcançar um determinado universo ao ajudar a localizar direções, a avaliar situações ou a planejar recursos. Outra técnica pode ser a presença de gráfico que auxiliam na melhor visualização dos resultados. O Gráfico 1, por exemplo, representa uma possível evolução da associação existente entre Vitor e Laura, em escala mensal.

Gráfico 1 - Evolução temporal do relacionamento entre Vitor e Laura.

Fonte: Autor.

Através desta técnica acompanha-se o relacionamento implícito entre as entidades em decorrência do avanço temporal. Com o passar do tempo às entidades estão cada vez mais associadas, sendo que neste momento, quando o gráfico atinge 0,9 pode ser sinalizada a aproximação entre as mesmas, antes mesmo de ocorrer à correlação. Está aproximação poderá resultar em um período relativamente curto de tempo, o mencionamento das entidades conjuntamente.

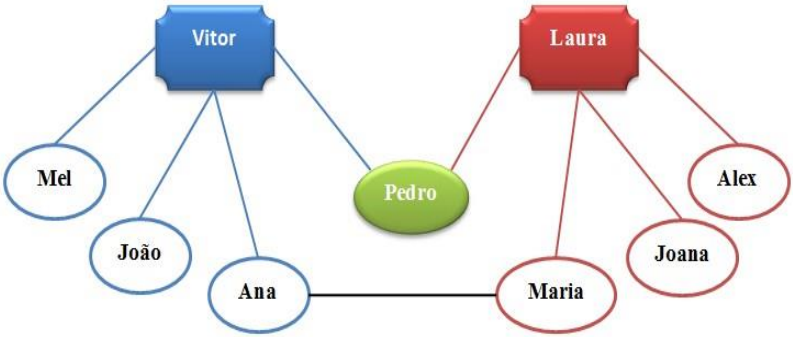
Existem também outras formas de representar a informação obtida. Outra maneira é através do mapa de tópicos. Na visão de Oliveira, Marcos, Vaasan (2000), um mapa tem em si vários tópicos que são ligados por devidas associações e subjacentes características. Para Biezunski e Newcomb (2001), os mapas de tópicos servem de representação da informação para descrever e navegar pela mesma. Tópicos podem ter vários nomes e ocorrências, e os escopos definem os limites de validade dos nomes, ocorrências e associações (PARK, 2003).

Segundo Silva (2008), um mapa de tópicos requer que o autor do mesmo pense em termos de tópicos (assuntos, ideias, conceitos), e associe várias informações aos mesmos. No início estes mapas eram usados para criar índices e glossários para documentos, posteriormente estendeu-se a internet (LIMA; FAGUNDES, 2004). Segundo Ahmed e Moore (2006), uma das principais vantagens apresentadas pelos mapas de tópicos é a questão da separação entre a estrutura conceitual e os recursos indexados, com isso obtêm-se grandes benefícios como:

- Representação das informações indexadas em mais alto nível, o que facilita a localização dos recursos envolvidos;
- Facilidade na hora da fragmentação, caso seja necessário a cerca de atender diferentes necessidades;
- Facilidade de combinação entre mapas de tópicos que indexam diferentes conjuntos de recursos.

A Figura 6 representa o mapa de tópicos aplicado ao exemplo dado sobre a diferença entre correlação e associação.

Figura 6 - Representação do mapa de tópicos.



Fonte: Autor.

Neste trabalho utiliza-se como técnica de explanação de relacionamentos indiretos a utilização de gráficos e mapa de tópicos.

2.5.1 Modelo Vetorial

O modelo vetorial visa estabelecer a similaridade entre os termos constantes em determinado vetor. O vetor é representado por uma entidade em questão e os demais termos correlacionados à mesma, trazendo o grau de correlação existente entre eles. Sendo assim, a partir de determinado termo (origem) é obtido um vetor de contexto que contém os termos coocorrentes. Na Tabela 3 podem ser observados os vetores de contexto de Vitor e Laura, entidades em questão.

A partir de uma equação de similaridade é possível chegar ao grau de semelhança entre dois conjuntos, representado por um número positivo (EGGHE, MICHEL; 2002). Jones e Furnas (1987) afirmam que a aplicação da equação de similaridade ocorre a partir de um par de vetores. Com este pressuposto pode-se utilizar equações de similaridade com o objetivo de extrair o grau de semelhança entre os vetores de contexto. Dentre diversas equações de similaridade destacam-se: o índice Jaccard, índice Dice, medida overlap (máxima e mínima), medida do cosseno e medida do pseudo-cosseno (EGGHE; MICHEL, 2002; JONES; FURNAS, 1987).

Segundo Gonçalves (2006) "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". Esta equação pode ser definida como:

$$\cos\theta = \frac{\sum_{i=1}^{n}(t1_i \times t2_i)}{\sqrt{\sum_{k=1}^{n}(t1_k)^2} \times \sqrt{\sum_{j=1}^{n}(t2_j)^2}}$$

onde t1 e t2 representam os vetores de contexto, $t1_i$, $t2_i$, $t1_k$, e $t2_j$ representam a frequência individual ou o peso dos termos t1 e t2.

O ângulo é gerado através da aplicação da equação do cosseno a partir dos vetores de contexto em questão, no caso "Vitor" e "Laura" representados na Tabela 3. O cálculo pode ser observado a seguir:

$$\cos(\overrightarrow{Vitor}, \overrightarrow{Laura}) = \frac{\sum_{i=1}^{n}(t1_i \times t2_i)}{\sqrt{\sum_{k=1}^{n}(t1_k)^2} \times \sqrt{\sum_{j=1}^{n}(t1_j)^2}}$$

$$\sum_{i=1}^{n}(t1_i \times t2_i) = (0.1 \times 0) + (0.2 \times 0) + (0.1 \times 0) + (0.4 \times 0.5) + (0 \times 0.2) + (0 \times 0.1) + (0 \times 0.3) = 0.2$$

$$\sqrt{\sum_{k=1}^{n}(t1_k)^2} = 0.1^2 + 0.2^2 + 0.1^2 + 0.4^2 = \sqrt{0.22} \cong 0.47$$

$$\sqrt{\sum_{j=1}^{n}(t1_j)^2} = 0.5^2 + 0.2^2 + 0.1^2 + 0.3^2 = \sqrt{0.39} \cong 0.62$$

$$\cos(\overrightarrow{Vitor}, \overrightarrow{Laura}) = \frac{0.2}{0.47 \times 0.62} \cong 0.68$$

Com o cálculo referente ao modelo vetorial observa-se que as entidades em questão, Vitor e Laura, possuem um percentual de aproximadamente 68% de similaridade, refletindo um grau elevado de semelhança.

3 COMPUTAÇÃO DISTRIBUÍDA

A evolução dos computadores representou um grande passo para as ciências exatas. As ideias para automatizar os cálculos e os períodos de guerra contribuíram de maneira expressiva para os avanços e invenções na área tecnológica. Há cada período, novos computadores surgem melhores e com maiores capacidades de processamento.

O breve contexto histórico sobre a evolução dos computadores que será feita, foi embasada nos estudo de Kowaltowski (1996) e Fonseca (2007). Na antiguidade foram dados os primeiros passos a fim de conseguirem-se formas de realizar cálculos de maneira automatizada, utilizando pedras e outros dispositivos que deram origem aos ábacos. Desde então muitos são os esforços para alcançar máquinas capazes de processar grandes volumes de dados. Leonardo Da Vince foi o responsável pela ideia de uma somadora mecânica. Em meados de 1600 surge então a pascalina, primeira somadora mecânica, desenvolvida por Blaise Pascal. Em 1801, Joseph Marie Jacquard inventa um tear mecânico, com uma leitora automática de cartões. O sucesso foi tanto que em torno de sete anos decorridos desde a invenção já havia 11 mil teares desse tipo operando na França.

O século XVIII é responsável por grandes avanços representados, por exemplo, por Charles Babbage que desenvolveu a máquina de diferenças e a máquina analítica e por Hollerit responsável por uma perfuradora e tabuladora de cartões. Já no século XIX mais avanços surgem com Alan Turing que desenvolveu a ideia da máquina universal capaz de executar qualquer algoritmo, formando assim, a base da computação. Em 1943 em um projeto coordenado pelo próprio Alan Turing, foi desenvolvido o *Colossus*, um computador inglês utilizado na segunda guerra mundial. Alan Turing foi ainda responsável por grandes avanços na área criptográfica, e juntamente com a equipe de *Bletchley Park* na Inglaterra quebraram os códigos cifrados da ENIGMA (máquina de cifrar usada pelos alemães), fato marcante para a computação quântica. Em 1944 contribuiu de forma direta no projeto de fabricação de computadores, assessorando a Eckert e John Machly, criadores do *ENIAC* em 1946, e que mais tarde construiriam o *UNIVAC* em 1950. Durante 1936 e 1939 o engenheiro alemão Konrad Zuse construiu o primeiro computador eletromecânico binário programável, o qual fazia uso de relés elétricos para automatizar os processos. John V. Atanasoff tem o crédito da patente do primeiro computador digital (1939).

Em virtude do aumento da necessidade de processar vasta quantidade de informação é necessário que as máquinas possuam um hardware com alto poder de processamento, para realizar as inúmeras atividades solicitadas pelo usuário. Gordon Moore foi um dos primeiros a expressar tal preocupação, em seu artigo "*Cramming more components onto integrated circuit*[2]". Moore afirmou que desde a década de 50, a prática de tornar os componentes eletrônicos cada vez menores, para a inclusão de funcionalidades mais complexas em um espaço cada vez mais reduzido, vem sendo difundida entre todos os pesquisadores desde a sua época. Dessa forma, a cada 18 meses[3] o nível de transistores reunidos em um mesmo circuito integrado dobraria, ou seja, o nível de complexidade aumentaria nos dispositivos (MOORE, 1965). Esse documento ficou conhecido como a Lei de Moore.

A evolução proposta por Moore segue ainda hoje a estimativa proposta por ele na década de 70, denominada lei de Moore, que prevê a duplicação do número de transistores comportados em uma pastilha a cada 18 meses (MOLLICK, 2006). Porém se for analisado o paradigma proposto por Moore quanto ao nível de processamento ele foi fragmentado recentemente em decorrência do avanço dos componentes e a criação de ambientes paralelos e distribuídos, que puderam fornecer um alto desempenho se comparado aos sistemas centralizados.

De 1945 a 1985 os computadores eram máquinas grandes e caríssimas sendo em sua maioria independentes devido a não existência de uma forma confiável de interligá-los (TANENBAUM, 1992). Surgiram, desde então, redes que permitem a conexão de vários computadores e a transmissão de dados a uma alta velocidade.

Para Foster (2002), um computador pessoal no ano de 2001 era tão rápido quanto um supercomputador em 1990. Em 1990 a capacidade de armazenamento de um supercomputador representava 100 GB, mesma capacidade em 2001 de um computador pessoal.

O modelo distribuído representa inúmeras vantagens para as organizações. Silberschatz, Galvin e Gagne (2010), destacam que as vantagens deste modelo têm seus resultados aliados a uma nova tendência da indústria em direção à redução do tamanho dos computadores. Ressaltam ainda que muitas empresas têm substituído computadores de grande porte, os mainframes, por redes formadas por computadores pessoais, trazendo vantagens empresariais como: relação custo-benefício, maior flexibilidade na alocação de recursos e expansão

[2] Em português: Agrupando mais componentes em circuitos integrados.
[3] Moore atualizou está estimativa para 2 anos.

das instalações, interfaces melhores para usuários e maior facilidade na hora de realizar as manutenções caracterizando-se assim por um modelo de computação distribuída.

A computação distribuída caracteriza-se por um sistema onde os componentes de hardware ou software localizados em computadores se comunicam e coordenam suas ações apenas trocando mensagens entre si, quando interligados em uma rede de computadores, permitindo o compartilhamento de recursos e informações entre os mesmos (COULOURIS; DOLLIMORE; KINDBERG, 2005). Para Tanenbaum e Steen (2007), a mesma é caracterizada por conjunto de computadores independentes que se apresenta a seus usuários como um sistema único e coerente. Silberschatz, Galvin e Gagne (2010), ressaltam que um sistema distribuído corresponde a uma coleção de computadores que não compartilham memória ou relógio, ou seja, cada processador possui sua própria memória local.

A computação distribuída pode ser vista também como um processamento distribuído em larga escala, ou processamento de alto desempenho, com o intuito de agregar recursos computacionais dispersos localmente ou geograficamente, sendo estes componentes: de hardware, pacotes de softwares, incluindo até instrumentos geograficamente dispersos (DANTAS, 2005). Os instrumentos dispersos podem ser telescópios ou aceleradores de partículas, como o LHC (Grande Colisor de Hádrons) que pode vir a produzir 15 petabytes ao ano, de acordo com o site da Organização Europeia de Pesquisa Nuclear (CERN)[4].

A computação distribuída requer não somente um único computador com alto poder de processamento, como por exemplo, mainframes, para executar suas operações, e sim a utilização de vários núcleos de processamento com a finalidade de distribuir a carga de processamento. A mesma pode ser vista através da seguinte frase "Dividir para conquistar". Em decorrências das evoluções tecnológicas, hoje já é possível o auxílio dos sistemas distribuídos até na medicina.

Afim de uma maior compreensão sobre os sistemas distribuídos a seguinte analogia pode ser feita: Ana vai até o supermercado para realizar as compras do mês. Após longas horas no estabelecimento, Ana conclui que terminou suas compras. Ao dirigir-se ao único caixa, Ana se depara com uma enorme fila e passa várias horas esperando até que possa enfim efetuar a compra. Este modelo remete ao trabalho de um único caixa de supermercado que pode ser comparado a único

[4] http://public.web.cern.ch/public/en/LHC/Computing-en.html

computador efetuando o processamento. Mudando a situação, Ana termina suas compras e dirige-se para efetuar o pagamento, ao chegar próximo ao caixa, Ana se depara com vários caixas de supermercado, podendo escolher a melhor fila, para assim pagar suas compras. Este modelo nos remete a computação distribuída, ou seja, vários núcleos de processamento.

De modo geral, uma arquitetura distribuída pode ser decomposta essencialmente em três categorias: multiprocessador, multicomputador e sistemas distribuídos. Na Tabela 5 verifica-se algumas características dos modelos mencionados.

Tabela 5-Características de Multiprocessador, Multicomputador e Sistemas Distribuídos.

Item	Multiprocessador	Multicomputador	Sistema Distribuído
Configuração do nó	CPU	CPU, RAM, Interface de rede	Computador completo
Periféricos do nó	Tudo compartilhado	Exc. Compartilhada, talvez disco	Conjunto completo por nó
Localização	Mesmo rack	Mesma sala	Em locais diferentes
Comunicação entre nó	RAM compartilhada	Interconexão dedicada	Rede tradicional
Sistemas operacionais	Um compartilhado	Múltiplos, mesmo	Possivelmente todos diferentes
Sistemas de arquivos	Um compartilhado	Um compartilhado	Cada nó tem seu próprio
Administração	Um compartilhado	Um compartilhado	Várias organizações

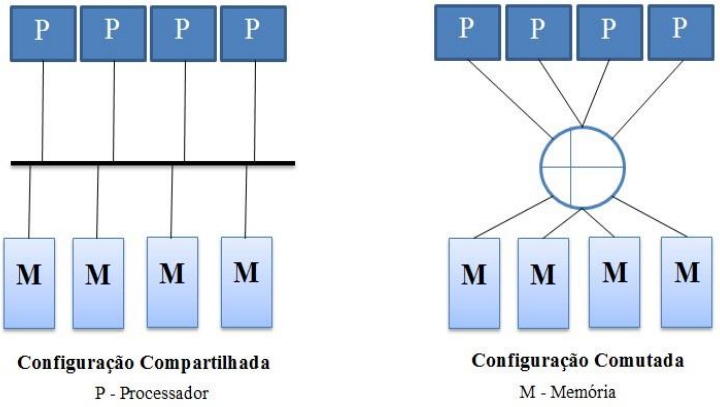
Fonte: (TANENBAUM; STEEN, 2007).

Os sistemas com múltiplos processadores possuem duas ou mais unidades de processamento trabalhando em conjunto, comunicando-se por um meio físico (TANENBAUM, 2010). Esta característica

possibilita a execução paralela de instruções, oferecendo um aumento no desempenho quando comparado ao processamento sequencial.

Para Tanenbaum (2010), um sistema de computador fortemente acoplado no qual duas ou mais unidade de processamento compartilham acesso total a memória pode ser visto como um Multiprocessador. Os processos utilizam a memória para realizarem a comunicação entre si, o procedimento de escrita pode ser visto pelos demais. Devido ao compartilhamento de memórias existente, este modelo utiliza mecanismos de sincronização de processos com objetivo de controlar a ordem de acesso à mesma. Como exemplifica a

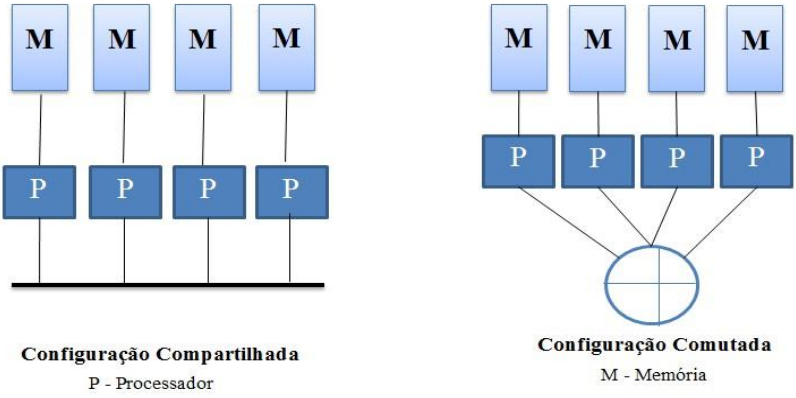
Figura 7 – Configurações genéricas de multiprocessadores.



Fonte: Adaptado de (DANTAS, 2005).

No âmbito dos multicomputadores, estes podem ser entendidos como um computador formado por várias unidades de processamento, ao qual não ocorre o compartilhamento da memória entre as mesmas, fazendo com que cada unidade de processamento possua uma memória local. De acordo com Tanenbaum (2010), multicomputadores são fortemente acoplados. A comunicação entre processos é feita pela troca de mensagens entre os processos em execução. Os nós de uma estrutura de multicomputador geralmente possuem CPU, RAM, uma interface de rede e talvez um disco rígido para a paginação. O meio físico de comunicação entre os nós do multicomputador é de alta velocidade, tornando sua escalabilidade reduzida se comparado a um sistema distribuído e geralmente trabalham paralelamente. A Figura 8 representa as configurações de multicomputadores

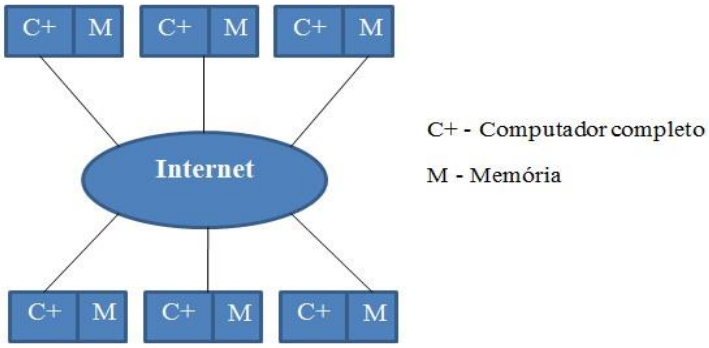
Figura 8 – Configurações genéricas de multicomputadores.



Fonte: Adaptado de (DANTAS, 2005).

Um sistema distribuído é semelhante a um multicomputador, porém podem ser interligados por redes normais (ETHERNET). Em sistemas distribuídos, computadores remotos cooperam via rede a fim de parecerem uma máquina local (DEITEL; DEITEL; CHOFFNES, 2005). Os autores ressaltam que os sistemas distribuídos surgem principalmente da necessidade de melhorar a capacidade, seja ela no âmbito do processamento ou do armazenamento, e a confiabilidade de uma única máquina. Geralmente nesta estrutura cada componente da estrutura é um sistema completo, com todos os periféricos, e executam os processos de forma distribuída, como exemplifica a Figura 9.

Figura 9 – Configurações genéricas de sistemas distribuídos.



Fonte: Adaptado de (TANENBAUM, 2010).

3.1 OS *MIDDLEWARES*

Os *middlewares* são responsáveis por fazer o interfaceamento entre a aplicação e o usuário. Permitem a um sistema distribuído manter-se uniforme mesmo operando em *hardwares* e sistemas operacionais distintos (TANENBAUM, 2010). Possui a capacidade de adaptarem-se dinamicamente às possíveis variações do ambiente e dos requisitos que os sistemas exigem (SILVA JÚNIOR, 2008).

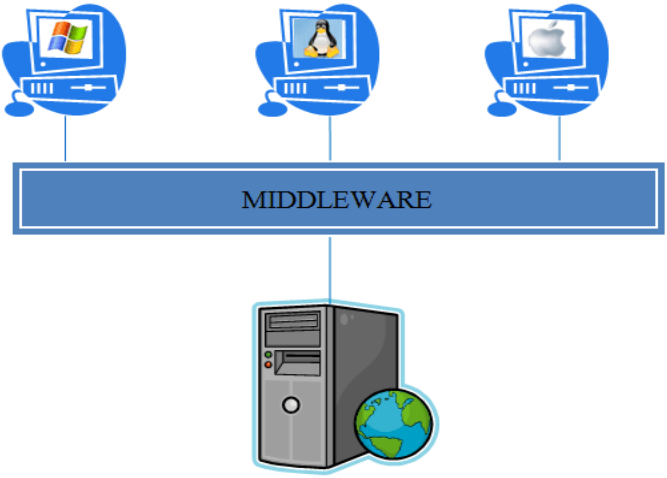
Tais serviços propiciam à aplicação simplicidade e transparência no uso e compartilhamento de recursos, ou seja, para a aplicação é como se houvesse um único processador (VIEIRA JÚNIOR, 2009; BERNSTEIN, 1996). Fornecem uma visão uniforme de redes, protocolos, e recursos de sistemas operacionais heterogêneos (SILVA JÚNIOR, 2008). Berman, Fox e Hey (2003) afirmam que o *middleware* "mascara" a complexidade da infraestrutura, simplificando a comunicação da camada de aplicação com a camada de recursos.

Dentre as formas de transparência ofertadas pelos *middlewares*, segundo Tanenbaum (2010) e Deitel, Deitel e Choffnes (2005) pode-se citar:

- Acesso: oculta diferenças na representação de dados e no modo de acesso a um recurso.
- Localização: oculta o lugar em que um recurso está localizado.
- Migração: Oculta que um recurso pode ser movido para outra localização.
- Relocação: oculta que um recurso pode ser movido para uma outra localização enquanto em uso.
- Replicação: oculta que um recurso é replicado em diversas máquinas.
- Concorrência: oculta que um recurso pode ser compartilhado por diversos usuários concorrentes.
- Falha: oculta a falha e a recuperação de um recurso.

A Figura 10 permite perceber um ambiente heterogêneo, ou seja, com várias máquinas com características técnicas diferentes (sistemas operacionais diferentes), executando em um ambiente distribuído interfaceado por uma camada (*middleware*).

Figura 10 – Ambiente interfaceado pelo *middleware*[5].



Fonte: Autor.

3.2 CARACTERÍSTICAS DE SISTEMAS DISTRIBUÍDOS

A computação distribuída consiste em um conjunto de máquinas conectadas por uma rede de comunicação, atuando como um sistema único. Esse sistema, ao contrário dos sistemas paralelos, possuem seus recursos próprios, como: memória principal e *clock* do processador. A comunicação é estabelecida por uma rede através de protocolos específicos (FOSTER; KESSELMAN; TUECKE, 2001). Além disso, o sistema distribuído precisa ser transparente, tolerante a falhas, escalável e ter um alto poder de processamento.

Segundo Deitel, Deitel e Choffnes (2005); Silberschatz, Galvin e Gagne (2010), pode-se citar as seguintes características básicas em relação aos sistemas distribuídos:

- Confiabilidade: o objetivo central do desenvolvimento de sistemas distribuídos é torná-los mais confiáveis que os sistemas centralizados, buscando uma utilização plena. Para alcançar uma disponibilidade plena é necessário aumentar o número de cópias de

[5]As imagens apresentadas na figura 10 se referem respectivamente aos sistemas operacionais Windows ™, Linux ™ e Mac OS™.

certas peças-chave e algumas funções importantes, ou seja, aumentar a redundância. A disponibilização de maiores quantidade de dados poderá implicar em uma inconsistência alta, isto devido ao fato que é difícil controlar todas as cópias. Mesmo mantendo a integridade entre elas, acaba-se por esbarrar em outro problema, a queda de desempenho por haver repetições. Outro aspecto relacionado à confiabilidade é a tolerância a falhas em que o sistema não deve parar, ele apenas mascara as falhas ocorridas, ou seja, caso ocorra uma falha os outros nós podem assumir e continuar a operar.

- Escalabilidade: soluções desenvolvidas para sistemas distribuídos de pequeno porte não funcionam adequadamente para os sistemas distribuídos de grande porte. Por exemplo, normalmente em sistemas distribuídos de pequeno porte, uma tabela de e-mails poderia ser armazenada em apenas um servidor, enquanto que em um sistema de grande porte isso não deve ocorrer. Isto devido ao fato de o sistema ficar a mercê do funcionamento deste servidor, uma vez que todas as aplicações que necessitam utilizar essa tabela só funcionam se o servidor estiver funcionando.
- Desempenho: é necessário que a taxa de execução do sistema seja alta. A aplicação executada em um ambiente distribuído não deve ser mais lenta que a executada em um ambiente centralizado. Para acompanhar o desempenho do sistema, métricas como o tempo de resposta e o número de *Jobs* por hora (*throughput*) são utilizadas, sendo assim, estas variáveis ficam condicionadas ao desempenho da rede.
- Transparência: característica relevante dos sistemas distribuídos. Para chegar ao nível de transparência têm-se duas situações sendo elas no nível do usuário, ou seja, o usuário interage com uma interface, mas não sabe onde ela está executando ou armazenando seus arquivos, ou no nível dos programas, sendo que neste caso o ambiente distribuído pode aumentar ou diminuir, transparente para a aplicação.
- Comunicação: um sistema distribuído pode proporcionar a execução de funções em grandes distâncias. Duas pessoas localizadas em regiões diferentes, ou até países diferentes, podem colaborar em um mesmo projeto.

Os sistemas distribuídos podem ser classificados quanto à sua organização e a finalidade com que são utilizados.

3.2.1 *Cluster*

Os *clusters* são computadores fracamente acoplados, criados para obter maior desempenho ou disponibilidade. Geralmente possuem um alto poder de processamento. Esta estrutura comporta-se como um sistema único composto por um conjunto de dispositivos idênticos (tanto pelo *hardware* quanto pelo *software* utilizado), visando o alto desempenho, o que confere um baixo custo devido à utilização de máquinas. Na maioria dos casos a computação de *cluster* é utilizada para programação paralela, na qual um único programa é executado paralelamente em vários computadores (TANENBAUM; STEEN, 2007). A principal característica desse modelo é a sua homogeneidade entre os dispositivos, ou seja, em grande parte possuem o mesmo sistema operacional e estão conectados a mesma rede.

Tanenbaum e Steen (2007), afirmam "que em um *cluster* o *hardware* subjacente consiste em um conjunto de estações de trabalho ou computadores pessoais semelhantes, conectados por meio de uma rede local de alta velocidade". Quanto às características deste modelo Stallings (2010), afirma que: "Um *cluster* consiste em um conjunto de computadores completos, conectados entre si, que trabalham juntos como um recurso computacional unificado, criando a ilusão de ser uma única máquina".

Como referido por Baker, Buyya e Hyde (1999), é aceito que um *cluster* de estações de trabalho (*workstations*) de alta performance consegue competir com os melhores supercomputadores que a IBM ou a SGI oferecem, sendo que uma organização pode montar um *cluster* deste tipo por cerca de $50000 enquanto a construção de um supercomputador custaria cerca de $200000, ou seja, quatro vezes mais. Ainda segundo Baker, Buyya e Hyde um Cluster é capaz de oferecer maior confiabilidade e segurança que os supercomputadores e, se for bem concebido, ainda maior tolerância a falhas.

3.2.2 *Grid*

Na computação em grade (*grid*), o ambiente é baseado na dispersão das funcionalidades de cada servidor. A especialização é uma das principais características deste modelo o que proporciona maior segurança aos serviços. Este tipo de modelo pode ser estruturado para que vários dispositivos atuem em conjunto, constituindo uma única aplicação. Cada máquina ficará encarregada de fornecer um

determinado serviço. O principal objetivo deste ambiente é garantir a disponibilidade e confiabilidade dos serviços. Ao ocorrer uma falha em determinado local, o sistema não cairá e continuará sua execução normalmente.

Buyya e Venugopal (2005), afirmam que um *grid* computacional é uma infraestrutura que envolve o uso integrado e colaborativo de computadores, redes, bancos de dados e instrumentos científicos que podem pertencerem e serem geridos por várias organizações. Também é possível afirmar que "... cada sistema pode cair sob um domínio administrativo diferente, e podem ser muito diferente no que tange a hardware, software e tecnologia de rede empregada" (TANENBAUM; STEEN, 2007).

Os *grids* possuem uma habilidade de compartilhar, selecionar e agregar recursos distribuídos geograficamente para solucionar problemas de larga escala na ciência, engenharia e comércio (CHETTY; BUYYA, 2002). Berman, Fox e Hey (2003), destacam que os grids permitem o gerenciamento ao acesso dos recursos remotos de maneira escalável, seguro e com alto desempenho, fornecendo um cenário onde ocorre a colaboração científica visando o objeto de compartilhar recursos para grupos de trabalhos distribuídos geograficamente.

Outra característica importante que permite a *grid* ser altamente escalável é a possibilidade de utilizar recursos heterogêneos, ou seja, ao optar por um ambiente que funcionará através do modelo em *grid* nenhuma premissa pode ser adotada em relação a *hardwares*, a sistemas operacionais, redes, domínios administrativos, politicas de segurança, entre outras (TANENBAUM; STEEN, 2007). É válido afirmar que esta estrutura computacional também pode possuir *clusters* em sua composição.

A ideia de terem-se *grids* computacionais é aproveitar o tempo ocioso que o computador possui, sendo eles recursos de diversos domínios, quando conectado a rede e utilizá-lo em prol do processamento. Em geral os recursos de diversas organizações são reunidos para permitir a colaboração entre grupos de pessoas, pesquisadores, instituições, enfim, formando uma organização virtual, conectadas via rede. Os recursos do ambiente que formam a *grid* não são dedicados, tornando possível sua utilização mesmo que em outrora sejam dedicados a outros fins, como computadores de empresas, universidades, de uso pessoal, entre outros. Recursos podem ser sensores, dados, computadores, etc. Através disto a arquitetura de um grid deste modelo pode ser verificada na Figura 11 que mostra ainda a arquitetura do protocolo da internet.

Figura 11 – Arquitetura em camadas.

Fonte: Adaptado de (FOSTER; KESSELMAN; TUECKE, 2001).

As diversas camadas deste *grid* fornecem recursos diferentes. Tanenbaum e Steen (2007), complementam que:

- A camada de base é responsável por prover interfaces para recursos locais em um site específico;
- A camada de conectividade consiste em protocolos de comunicação para suportar transações da grade que abranjam a utilização de múltiplos recursos;
- A camada mais acima, a de recursos é responsável pelo gerenciamento de um único recurso;
- A coletiva provê a manipulação do acesso a múltiplos recursos e normalmente consiste em serviços para descoberta de recursos, alocação e escalonamento de tarefas para múltiplos recursos, replicação de dados, dentre outros.

3.2.3 Computação em Nuvens

O termo Computação em Nuvens é também conhecido como *Cloud Computing*, e vem sendo difundido nas últimas décadas como um novo paradigma da computação distribuída. O modelo baseado em computação em nuvens vale-se da computação em *grid* para trabalhar

com a computação distribuída, porém os dados gerados ficam difundidos em diferentes locais, influenciando a ideia de que os mesmos encontram-se em nuvens computacionais.

Segundo Taurion (2009), o termo *cloud computing* ou computação em nuvem foi mencionado primeiramente em uma palestra no ano de 2006 por Eric Schmidt, funcionário do Google, sobre como sua empresa gerenciava seus *data centers* (centro de processamento de dados).

Para Armbrust et al. (2009), a computação em nuvens é um conjunto de serviços de rede ativados, proporcionando escalabilidade, qualidade de serviço, infraestrutura barata de computação sob demanda e que pode ser acessada de uma forma simples. Para Buyya, Broberg e Brandic (2009), o fato de utilizarmos o termo nuvem é apenas uma metáfora, pois ela representa a Internet ou infraestrutura de comunicação entre os componentes que envolvem a arquitetura, ou seja, abstraindo a complexidade. Os serviços pertencentes a esta infraestrutura advêm de serviços, normalmente alocados em centros de processamento de dados.

A infraestrutura que permite desenvolver o ambiente da computação em nuvem normalmente é composta por um número elevado de máquinas físicas conectadas através de uma rede. Cada máquina física tem as mesmas configurações de software, mas pode ter variação na capacidade de hardware em termos de CPU, memória e armazenamento em disco (SOROR et al, 2010). Na visão de Weiss (2007), a *Cloud Computing* é uma promessa de serviços confiáveis que serão disponibilizados a partir da próxima geração de *data centers*. Assim, os consumidores poderão acessar, de qualquer lugar do mundo, sejam elas aplicações ou dados, pois estão armazenados em uma nuvem computacional.

A evolução dos serviços e produtos no campo que abrange as tecnologias da informação acarretou em um novo modelo a computação em nuvem, também chamada de *Utility Computing* (BRANTNER et al. 2008). A *Utility Computing* tem como objetivo central fornecer os componentes básicos como armazenamento, processamento e largura de banda de uma rede como um "produto" através de provedores especializados com um baixo custo por cada unidade utilizada. O modelo baseado em *Utility Computing* não fornece preocupação quanto à escalabilidade, pois a capacidade de armazenamento fornecida é praticamente infinita.

A *Utility Computing* propõe fornecer disponibilidade total, isto é, os usuários podem ler e gravar dados a qualquer momento, sem nunca serem bloqueados. Os tempos de resposta são quase constantes e não

dependem do número de usuários simultâneos, do tamanho do banco de dados ou de qualquer parâmetro do sistema. Os usuários não precisam se preocupar com *backups* (cópias de segurança), pois se os componentes falharem, o provedor é responsável por substituí-los e tornar os dados disponíveis em tempo hábil por meio de réplicas (BRANTNER et al. 2008). A Figura 12 nos remete a estrutura que a computação em nuvens oferece.

Figura 12 – Representação dos dados na nuvem.

Fonte: Autor.

3.2.4 Vantagens dos sistemas distribuídos

Para Tanenbaum (1992), os sistemas distribuídos possuem vantagens sobre os sistemas centralizados devido aos seguintes fatores:

- Maior poder de processamento: um sistema distribuído pode ter um poder de processamento maior em relação aos mainframes;
- Crescimento Incremental: o poder computacional pode crescer incrementalmente;
- Compartilhamento de dados e recursos: algumas aplicações envolvem máquinas separadas geograficamente;
- Maior confiabilidade: caso uma máquina saia do ar, o sistema como um todo pode continuar executando.

- Menor custo/benefício: os sistemas distribuídos têm melhor custo/beneficio em relação aos sistemas centralizados, isto significa que o sistema distribuído é mais barato e oferece maior eficiência no processamento dos dados.

Complementando a ideia do autor pode-se citar ainda mais algumas vantagens como:

- Melhor desempenho: nos sistemas distribuídos vários processadores trabalham em conjunto, através deste fator chegam a superar grandes computadores como os mainframes.
- Maior carga de execução: com vários processadores em diferentes máquinas realizando o processamento, pode-se submeter uma carga maior de dados para serem processados.
- Maior número de usuários atendidos: com a capacidade de processamento maior, mais usuários podem ser atendidos. Pensando no âmbito do serviço bancário, com vastos processadores realizando as diversas operações solicitadas, mais usuários podem requerer ao serviço online.
- Melhor tempo de resposta: com a capacidade de processamento aumentada, pode-se executar mais rapidamente as operações.

3.2.5 Desvantagens dos sistemas distribuídos

Perante as diversas vantagens, as desvantagens não podem ser desconsideradas. Para Deitel, Deitel e Chouffnes (2005), os sistemas distribuídos podem ser difíceis de implementar e gerenciar. Um dos problemas existentes envolve a rede responsável por conectar o sistema distribuído, pois a tendência do mesmo é sempre crescer, sendo assim pode chegar um momento que a rede sature, o que provoca atrasos e problemas de confiabilidade introduzidos pelas redes subjacentes (DEITEL; DEITEL; CHOUFFNES, 2005). Apesar de a tecnologia existente permitir redes de altíssimas velocidades, isto seria comparável ao gargalo de Newman (a velocidade do processamento de máquina depende do barramento), ou seja, o sistema distribuído depende do desempenho da rede (TANENBAUM,1992). Outro ponto importante envolve o compartilhamento de dados. Os dados estão acessíveis a todos os que compõem a rede, com isso perde-se em termos segurança. Dados que necessitam ser de caráter secreto não podem ser acessíveis a todos,

mas estes dados estão disponíveis ao sistema, sendo assim e possível que o mesmo seja invadido sem autorização.

3.3 PROJETOS EXISTENTES DE SISTEMAS DISTRIBUÍDOS

O projeto *seti@home*[6] é um exemplo de *grid* computacional. SETI significa *Search for Extraterres Trial Intelligence* (busca por inteligência extraterrestre). Este projeto busca vida extraterrestre analisando sinais de rádio, através de aparelhos radiotelescópios, com objetivo de identificar mensagens emitidas por outras civilizações. O *SETI@HOME* foi lançado em 17 de maio de 1999. Em seus 10 anos de operação, atraiu mais de 5 milhões de participantes, localizados em 226 países (KORPEL et al., 2011).

Outro exemplo é o programa *folding@home*[7] da universidade de Stanford. O objetivo do projeto é estudar a estrutura das proteínas relacionadas com a cura de doenças como: Alzheimer, Câncer, Parkinson entre outras. O funcionamento é bastante similar ao SETI. O grupo do *Folding@home* se comprometeu a divulgar todas as descobertas desse projeto na Internet de forma livre, de modo que outros pesquisadores do mundo inteiro possam utilizar os dados em suas próprias pesquisas.

O projeto *Genome@home*[8] trabalha na teoria e simulações com proteínas, RNA e polímeros sintéticos em nano escala. O objetivo do projeto é a análise de proteínas e suas aplicações, o que teve implicações em muitas áreas, incluindo medicina. *Genome@home* foi executado pela *Pande Lab* da Universidade de Stanford, uma instituição sem fins lucrativos, dedicada à pesquisa científica e educação.

Einstein@Home[9] é um projeto baseado em computação distribuída executado sobre a plataforma de software *BOINC*. Ele procura por pulsares e ondas gravitacionais emitidas por pulsares, buracos negros, estrelas de nêutrons, estrelas de quarks e outros objetos bem densos, que, teoricamente, podem emitir fortes ondas gravitacionais.

[6] http://setiathome.ssl.berkeley.edu/
[7] http://folding.stanford.edu/
[8] http://genomeathome.stanford.edu/about.html
[9] http://www.physicscentral.com/experiment/einsteinathome/

World Community Grid[10] é um esforço para criar o maior supercomputador público do mundo para realizar pesquisas científicas que beneficiem a humanidade. O projeto é de autoria da IBM e atualmente está disponível para Windows, Linux, e Mac OS X. Utiliza a plataforma BOINC. O *World Community Grid* oferece múltiplos projetos humanitários para a participação utilizando um mesmo software. Os projetos são escolhidos criteriosamente por membros de grandes instituições de pesquisa e universidades do mundo inteiro.

Na Universidade Federal de Santa Catarina, mais precisamente no campus Araranguá, o projeto *AraBoinc*[11] caracterizou-se como um sistema distribuído, pois utilizou a capacidade de processamento das máquinas dos colaboradores associados ao projeto, quando ociosas. A ideia da competição foi popularizar e disponibilizar uma plataforma BOINC no campus Araranguá envolvendo para isto alunos, professores e técnicos. A capacidade ociosa do computador (ao acionar, por exemplo, a proteção de tela) foi direcionada ao processamento de simulações de sistemas físicos, e o competidor ganhava créditos por unidade processada. Ao final do período da competição, o vencedor foi aquele com maior número de créditos.

[10] http://www.worldcommunitygrid.org/
[11] http://www.araboinc.ufsc.br/boincufscara/

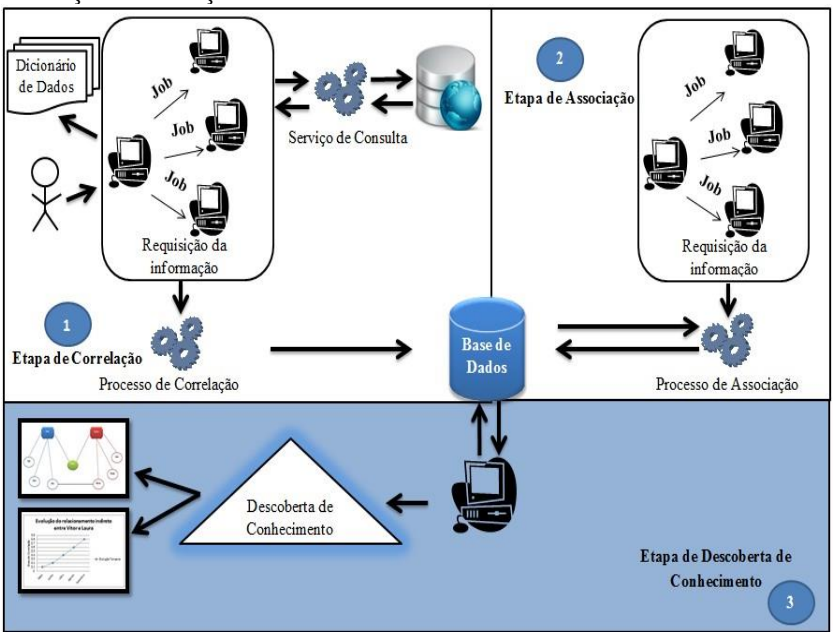
4 ARQUITETURA PROPOSTA

Neste capítulo será apresentada a arquitetura proposta. A apresentação dar-se-á em duas etapas, a primeira refere-se ao modelo lógico, sendo que o mesmo detalhará a interação decorrente entre os módulos componentes da proposição. A segunda etapa será responsável por apresentar o modelo físico, descrevendo os componentes tecnológicos, bem como, a justificativa da utilização dos mesmos.

4.1 MODELO LÓGICO

O modelo lógico representado pela Figura 13 é composto por uma série de etapas que possibilitam a interconexão do conteúdo textual, representado por conceitos em um domínio de problema, objetivando prover suporte a tarefa de descoberta de conhecimento.

Figura 13 – Modelo lógico de descoberta de conhecimento com base na correlação e associação de termos.



Fonte: Autor.

4.1.1 Etapa 1: Processo de Correlação

Esta etapa é responsável pelo processo de correlação entre os termos aos quais se deseja analisar em determinado domínio de problema. Para efetuar este processo é necessária a figura do especialista de domínio, onde o mesmo é responsável por fazer a manutenção da base de dados, ou seja, a inserção do dicionário de dados e acionar o início do processo de correlação. De modo geral, a etapa divide-se em Requisição da informação e Processo de Correlação.

Requisição da informação: neste passo são acessados os termos contidos no dicionário. O dicionário de dados contêm os termos (palavras-chave) aos quais será aplicada a pesquisa. Um termo é constituído por palavras que, quando associado a uma classe geram um conceito. A diferença entre termos e conceitos é que o segundo possui um significado, enquanto termos isolados representam apenas palavras que dependem da avaliação de um especialista para expressarem alguma semântica. Com o intuito de gerar conceitos, os termos são atribuídos de acordo com as classes pertencentes ao domínio de análise. Os dados ficam dispostos da seguinte forma:

- Classe: a classe agrega sentido ao termo, por exemplo, o termo Brasil, pertence à classe País, já Joana pertence à classe Pessoa.
- Domínio: o domínio pode ser entendido como o "domínio do problema". Desse modo, a inserção de conceitos em um domínio permite que os processos de correlação e associação sejam aplicados para um fim específico. Como exemplo de domínios podem ser citados tecnologia, agricultura, saúde, ou mesmo, um domínio genérico.
- Termo: palavra que se deseja pesquisar, por exemplo, crise, meio ambiente, entre outras.

Fica evidente que este processo é responsável por agregar sentindo a um termo.

Após o processo descrito anteriormente, no qual foram gerados conceitos, é necessário selecionar o domínio a qual os conceitos pertencem. O domínio pode ser entendido como o domínio do problema. Desse modo, a inserção de conceitos em um domínio permite que o processo de correlação seja aplicado para um fim específico.

De posse desses termos, o servidor então realiza requisições a uma base de dados por meio de um serviço web responsável por obter a quantidade de documentos em que o termo aparece. Essa quantidade é,

ao nível de projeto, interpretada como uma aproximação da frequência individual de determinado termo em uma base de dados específica que representa um domínio de problema em particular. O processo, ao ser finalizado, retorna os termos e suas respectivas frequências armazenando-os em uma base de dados.

Processo de Correlação: com as respectivas frequências individuais dos termos do dicionário de dados já contidas na base de dados, o processo de correlação é iniciado por meio da distribuição dos *Jobs*, ou seja, a carga de trabalho entre os computadores que compõem a *grid*, ao qual serão chamados de nodos da *grid*. A lista de termos do dicionário com as frequências individuais é repassada para cada *Job*, de modo que cada um possa realizar a combinação, busca conjunta, de um termo em particular com todos os demais termos.

Como analogia pode-se pensar no seguinte cenário em que se tenham dez nodos compondo a *grid* e vinte termos contidos no dicionário de dados. A divisão seria a seguinte: o primeiro nodo receberá o dicionário completo do nodo mestre, com os vinte termos e suas respectivas frequências individuais. O trabalho consiste na busca pela frequência conjunta do primeiro elemento do dicionário, com todos os demais. O segundo nodo realizará a execução com apenas 19 termos do dicionário de dados, ou seja, o primeiro será desconsiderado, em virtude de o primeiro nodo já ter realizado os cálculos referentes a ele. O segundo nodo então começa o processo de obtenção da frequência conjunta do segundo elemento do dicionário com todos os demais termos do dicionário. E assim dar-se-á a divisão da carga de trabalho entre nos nodos pertencentes à *grid*. Esse processo será repetido até obter-se a matriz de dados com suas respectivas frequências conjunta, conforme a Tabela 1 apresentada no Capítulo 2.

Cálculo referente ao coeficiente de correlação: com os valores da frequência individual e da frequência conjunta dos termos analisados já obtidas e devidamente armazenadas na base de dados, pode-se então calcular o coeficiente de correlação, o qual representará a força de correlação entre os dois termos. Algumas equações utilizadas para obter o coeficiente de correlação foram apresentadas no Capítulo 2, tópico referente a modelos baseados em coocorrência do presente trabalho.

4.1.2 Etapa 2: Início do processo de associação

Através da etapa 1 obtêm-se o valor do coeficiente de correlação entre os termos da pesquisa em questão. Para efetuar o processo de associação entre os termos é necessário acessar os valores correspondentes ao cálculo do processo anterior. Em virtude disto, os valores obtidos ficam armazenados em uma base de dados, sendo que a mesma é compartilhada entre o processo responsável pela correlação e pelo processo de associação. Após a conclusão do processo de correlação inicia-se então o processo de associação. Os nodos pertencentes à *grid* executarão o processo responsável por desvendar possíveis relacionamentos indiretos existentes entre termos contidos no dicionário de dados. Os nodos fazem uma consulta à base de dados para obter o vetor de contexto de cada um dos termos com seus respectivos coeficientes de correlação, ou seja, para cada termo existe um conjunto (vetor) de termos diretamente relacionados através da correlação. Após isso, executam o processo de associação, no qual foi apresentado no Capítulo 2, tópico referente ao modelo vetorial. Ao efetuar o cálculo do modelo vetorial, obtém-se o grau de associação entre os termos. Cabe ressaltar que o processo de associação é executado considerando dois termos que não possuem correlação direta, ou seja, que no domínio de análise, considerando determinado período, não foram mencionados conjuntamente. Este valor por sua vez ficará armazenado na base de dados.

4.1.3 Etapa 3: Processo de descoberta de conhecimento

Com o cálculo do processo de associação concluído pelos nodos da *grid*, obtém-se então o grau de similaridade entre os termos da pesquisa. O passo 3 é responsável pela descoberta de conhecimento considerando os dados gerados pelos processos de correlação e associação. A partir desse ponto a base de dados possibilita explorar seu conteúdo visando à obtenção de padrões e tendências que conduzam a descoberta de conhecimento relevante e útil à tomada de decisão. Isso pode ser obtido através de gráficos de correlação e associação, histogramas de termos com escala temporal, e mesmo, mapas de tópicos, temporais ou não.

4.1 MODELO FÍSICO

Nas próximas seções serão detalhados os componentes tecnológicos e os processos, e como estes se inter-relacionam visando oferecer uma visão integrada do modelo proposto.

4.2.1 Serviço de Consulta

O serviço de consulta utilizado foi um servidor de indexação que suporta o armazenamento de conteúdos textuais com diferentes contextos. O sistema foi desenvolvido por Conceição (2013). O objetivo principal do servidor é ser um sistema computacional que possibilitasse o armazenamento integrado de informação estruturada e não estruturada de diferentes domínios, permitindo posteriormente a disponibilização deste conteúdo. No momento do processo de indexação, com base em um documento específico é extraído o texto completo, e este por sua vez é usado para criar uma instância do mesmo em um formato apropriado, chamado de vetor do documento (CONCEIÇÃO, 2013). O conjunto desses vetores dá origem a uma estrutura baseada em índice invertido que promove suporte a consultas textuais.

4.2.2 Plataforma de apoio ao ambiente distribuído

Para realizar o processo de correlação proposto no modelo representado pela etapa 1 da Figura 13 e o processo de associação dos termos, etapa 2 da mesma figura, foi utilizado um *framework/middleware* para oferecer suporte ao desenvolvimento de uma aplicação distribuída. Um *framework*, segundo Mattsson (2000), é um *software* que trabalha com um conjunto de classes, que podem ser abstratas ou não, cujas instâncias trabalham em conjunto. Pode oferecer suporte a integração com outros *frameworks*, e o ideal é que seja reutilizável.

Para Resende (2010), a computação distribuída pode ser a solução ideal, pois através dela é possível utilizar a capacidade ociosa de vários computadores diferentes, separados fisicamente, ou até mesmo geograficamente, para executar tarefas ou armazenamento de dados.

O *framework/middleware* escolhido foi o *GridGain.* Para Resende (2010), o *GridGain* é um projeto que busca sanar pontos como a complexidade de implantação e a integração com projetos ou

aplicativos, provendo uma solução de fácil integração com projetos *Java*, abstraindo ao máximo a complexidade de implantação.

O *GridGain* é um software *middleware JVM-based (Java Virtual Machine - based)* que permite o desenvolvimento da computação intensiva de dados e aplicações distribuídas de alto desempenho. As aplicações desenvolvidas com o *GridGain* podem ser escaláveis em qualquer infraestrutura, desde um único dispositivo móvel até uma grande nuvem (IVANOV; DMITRIY, 2012).

O projeto responsável por manter o *GridGain* iniciou no ano de 2005, com o propósito de ser um projeto de código aberto. O primeiro lançamento oficial ocorreu no ano de 2007 e no final de 2010 já haviam conquistado o mercado. Atualmente, o projeto conta com a parceria de um grupo de empresas que se uniram para dar suporte ao desenvolvimento do software. O projeto possui duas versões: a *GridGain Community Edition* com código aberto, e a versão *GridGain Enterprise Edition*, que é a versão comercial do software. Neste trabalho a versão utilizada é a *Community Edition.*

4.2.3 Modelo Dimensional

A representação dos dados do modelo proposto utiliza o conceito *Data Warehouse* (DW). Segundo Inmon (1992) um DW é "uma coleção de dados orientada por assunto, integrada, não volátil, variante no tempo que dá apoio às decisões da administração". Os DW são orientados a assunto, pois o mesmo é organizado por assunto e não por aplicação, sendo assim ele contêm apenas as informações necessárias para o processamento (SINGH, 2001). Caracterizam-se também por serem integrados devido à maneira como armazenam os dados coletados de diferentes fontes em um formato consistente (GONÇALVES, 2003). Não voláteis, pois os usuários podem apenas realizar consultas sobre os dados e jamais alterá-los (GOUVEIA. 2009). Variáveis com o tempo, pois todos os dados coletados referem-se a um momento específico, o que permite traçar o histórico das alterações realizadas, pelo fato dos dados nunca serem sobrepostos (SANTOS, 2009).

Os *Data Warehouses* proporcionam uma análise mais complexa sobre os dados que serão acessados, voltados à tomada de decisão. Suportam assim grandes demandas por dados e informações (ELMASRI; NAVATHE, 2005). O mesmo corresponde ao processo de integração dos dados corporativos de uma empresa em um único repositório a partir do qual os usuários finais podem facilmente executar

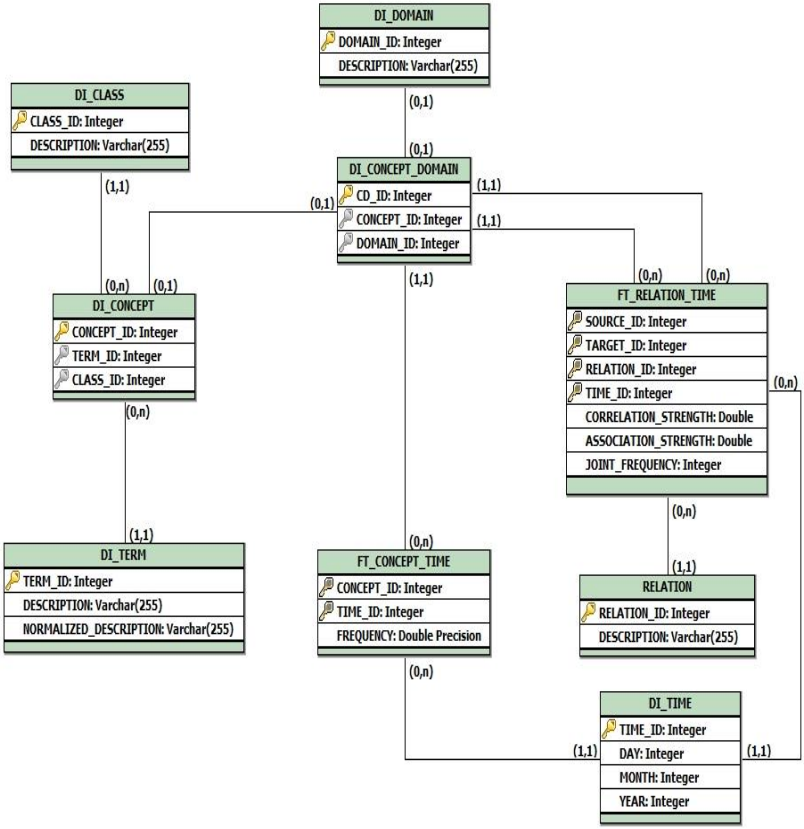
consultas, gerar relatórios e fazer análises (ELMASRI; NAVATHE, 2005; SINGH, 2001).

Para Gonçalves (2003), os DWs têm por objetivo disponibilizar os dados em uma modelagem facilmente entendível para os usuários finais. Segundo Singh (2001), "Métodos tradicionais de análise de dados, baseados principalmente no manuseio direto dos dados pelo homem, simplesmente não permitem a manipulação de conjuntos volumosos de dados".

A partir das vantagens citadas acima, é justificada a utilização de um modelo aplicando os conceitos de *Data Warehouse*, no modelo proposto, visto que este auxilia na tomada de decisão, através de análises complexas a partir de dados oriundos de bases textuais. A Figura 14 ilustra a modelagem de DW baseada em Silva (2012). O modelo foi adaptado para comportar o domínio, nos quais os campos SYNONYM_ID e VADID da tabela DI_CONCEPT, presentes na proposta de Silva, não foram utilizados.

É importante salientar que na modelagem dimensional existem formas específicas de realizar a mesma. Neste trabalho foi desenvolvida uma modelagem considerando os conceitos de tabelas de dimensões, onde são responsáveis por descrever determinado objeto (conceito) e tabelas de fatos, onde ficam localizadas as medidas, valores essenciais para a organização, assim como a chave primária das tabelas de dimensões. Como característica relevante deste modelo, tem-se a possibilidade de representar relacionamentos em escala temporal. Dentre as formas de realizar esta modelagem estão o modelo estrela e o modelo floco de neve, ao qual serão apresentados no próximo tópico.

Figura 14 – Modelo físico.



Fonte: Adaptado de (SILVA, 2012).

A seguir são detalhadas as tabelas (dimensões e fatos) que compõe o modelo do banco de dados:

DI_TERM: nesta dimensão são armazenados os termos sem qualquer contexto. Possui como atributos: (TERM_ID), (DESCRIPTION) e (NORMALIZED_DESCRIPTION). O primeiro corresponde a um valor sequencial, representando assim a identificação do termo, o segundo a uma breve descrição sobre o termo, o terceiro a uma descrição normalizada, ou seja, não contêm palavras pouco representativas (artigos, preposições, entre outras) e considera somente a raiz (sem sufixos) das demais palavras inseridas.

DI_CLASS: dimensão responsável por indicar a classe a qual pertence o termo. Muitas palavras ou termos podem representar um

duplo sentido, como por exemplo: manga, que pode ser pertencente à classe vestuário ou a classe fruta; ou até mesmo a palavra rosa, que pode ser pertencente à classe cor ou à classe planta. Em virtude disto, torna-se necessário indicar a classe ao qual pertence o termo. A tabela possui como atributos um número sequencial que identifica a classe (CLASS_ID) e outro indicando a descrição da classe (DESCRIPTION).

DI_CONCEPT: esta dimensão representa um conceito, sendo composta por um termo aliado a uma classe. A tabela é composta por um identificador do conceito ao qual pertence o termo (CONCEPT_ID); o identificador do termo (TERM_ID), como chave estrangeira, ou seja, que está vindo da tabela DI_TERM e o identificador da classe (CLASS_ID). Com as informações que se possui até agora, pode-se então refletir uma contextualização a cerca do termo.

DI_DOMAIN: esta tabela representa o domínio da análise. Um domínio é composto pelos seguintes atributos: identificador do domínio (DOMAIN_ID) e uma breve descrição acerca do mesmo (DESCRIPTION).

DI_CONCEPT_DOMAIN: esta dimensão é composta pelos atributos (DOMAIN_ID) identificador do domínio, (CONCEPT_ID) identificador do conceito, e (CD_ID) identificador sequencial que rotula o domínio e o conceito conjuntamente. A partir deste ponto já se possui um termo que estará associado a uma classe e a um domínio de pesquisa.

DI_TIME: por tratar-se de um protótipo que visa a implementação em escala temporal, faz-se necessária a dimensão DI_TIME. Esta tabela é responsável por armazenar a dimensão temporal, em que foram realizadas as consultas. Em virtude desta variável é necessário manter um identificador de tempo (TIME_ID), ou seja, um sequencial para controlar as diversas datas em que as consultas serão realizadas. Necessita-se ainda de mais três atributos, um para representar o dia (*DAY*), outro para representar o mês (*MONTH*) e outro para representar o ano (*YEAR*).

FT_CONCEPT_TIME: a tabela FT_CONCEPT_TIME é responsável por armazenar a frequência (FREQUENCY), o identificador do conceito (CONCEPT_ID), e um identificador temporal (TIME_ID), em virtude da necessidade da realização das consultas em diversos períodos de tempo, sendo que o mesmo identificador é representado como chave estrangeira, ou seja, está vindo da dimensão DI_TIME, onde ficarão guardados todos os dados referentes a tempo.

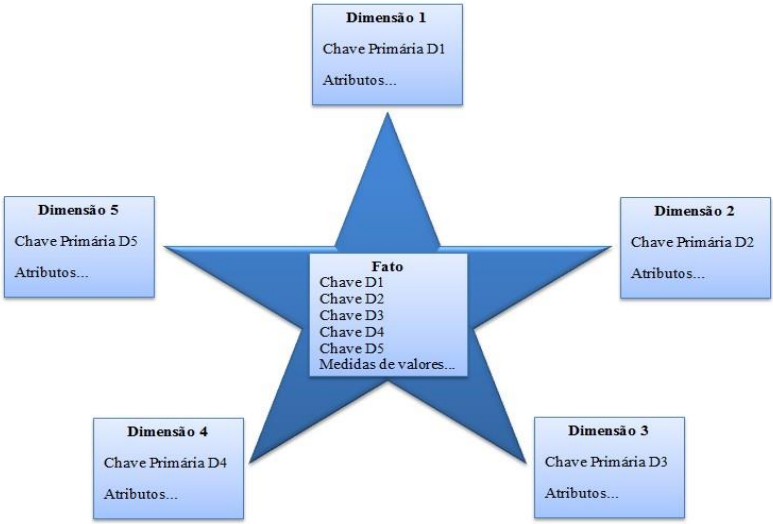
DI_RELATION: está dimensão é responsável por descrever (DESCRIPTION) a relação entre dois termos. Além da descrição há um campo identificador da relação (RELATION_ID).

FT_RELATION_TIME: a tabela FT_RELATION_TIME é responsável por armazenar os valores obtidos através da aplicação dos cálculos do modelo proposto. Dentre os valores obtidos encontram-se o valor de correlação (CORRELATION_STRENGTH) e o valor de associação (ASSOCIATION_STRENGTH). Estes valores são obtidos através da análise de dois termos (SOURCE_ID e TARGET_ID), os quais são ligados por algum tipo de relação (RELATION_ID). Para serem realizados os cálculos é necessário ainda armazenar a frequência conjunta entre os termos em uma determinada base de dados (JOINT_FREQUENCY). Para expressar a evolução temporal das consultar é armazenado o identificador da data (TIME_ID), ou seja, um identificador do tempo, armazenado na dimensão DI_TIME.

4.2.4 Modelo Estrela

O modelo estrela, ou *star schema*, representa uma forma de realizar a modelagem dimensional de maneira específica. Neste modelo temos a presença de dois pontos importantes que são: a) tabela de fatos que representa um acontecimento, ou seja, um fato ocorrido e os valores agregados destes fatos, também chamados de medidas; e b) tabelas de dimensões, que representam os dados qualitativos relacionados à tabela de fatos. A Figura 15 representa este modelo. No ponto central da figura encontra-se a tabela de fatos, nas extremidades as tabelas de dimensões. A tabela de fatos é onde as medidas numéricas do fato representado estão armazenadas.

Figura 15 – Representação do modelo estrela.

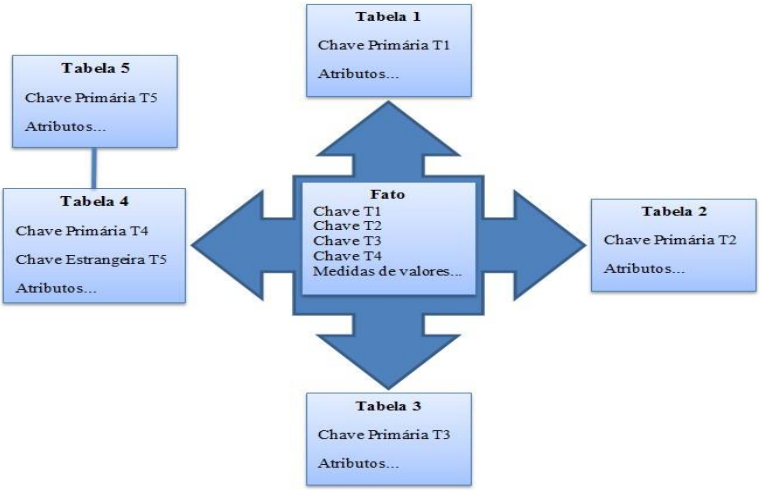


Fonte: Autor.

4.2.5 Modelo floco de neve

Outra forma de realizar a modelagem dimensional de maneira específica é o modelo floco de neve ou *snowflake*, que consiste em uma extensão do esquema estrela onde cada uma das "pontas" da estrela, representadas pelas tabelas de dimensões passa a ser o centro de outras estrelas. Isto porque cada tabela de dimensão seria normalizada, considerando a terceira forma normal, onde não deve ocorrer à dependência transitiva de chaves, e com isso geraria outras tabelas. Este modelo introduz tabelas dimensionais principais conectadas às tabelas de fato e tabelas dimensionais de extensão, onde são armazenadas as descrições das dimensões (MELLO, 2002). Estas tabelas de extensão são obtidas através da normalização das dimensões. A Figura 16 representa o modelo floco de neve onde a tabela 5 seria uma extensão da tabela 4, ou seja, a tabela 4 foi normalizada e passou a ser o centro de outra estrela, no caso, ligada a tabela 5.

Figura 16 – Representação do modelo floco de neve.

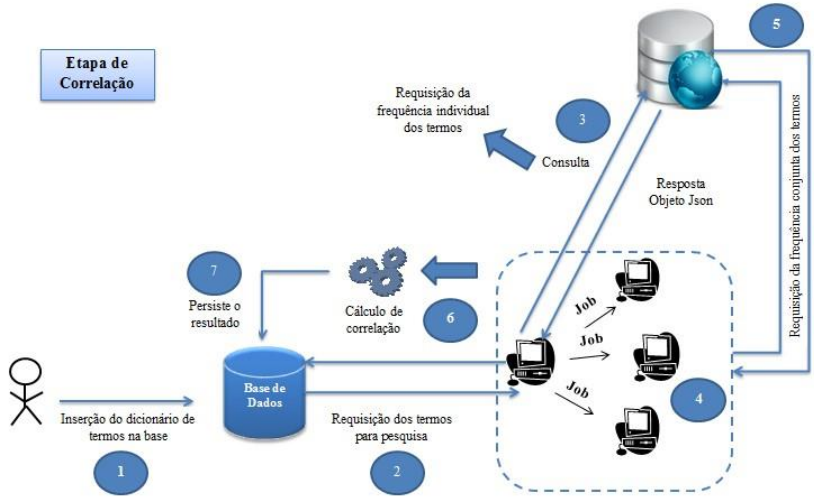


Fonte: Autor.

4.2.6 Detalhamento do processo de correlação

A seguir o serviço de correlação será descrito de maneira detalhada a partir da Figura 17 que representa a arquitetura física.

Figura 17 – Processo responsável pela correlação.



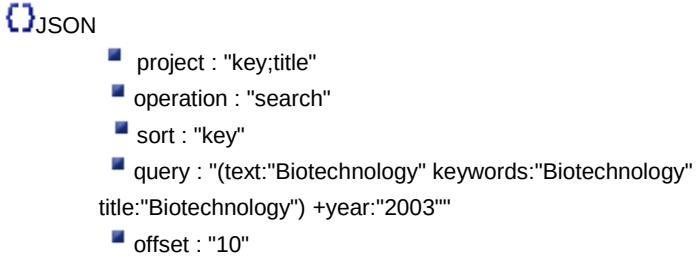
Fonte: Autor.

No passo 1 observa-se que ocorre a inserção do dicionário de dados. Esse passo caracteriza-se pela presença de um especialista de domínio, responsável por inserir na base de dados os termos aos quais se deseja pesquisar e aplicar o processo de correlação, e posteriormente o processo de associação. Neste momento é feita a leitura via aplicação, de todas as palavras chaves dos documentos, e estas são inseridas no banco de dados.

No passo 2 a comunicação entre a aplicação e a base de dados é realizada através da API JDBC (*Java Database Connectivity*). A aplicação que iniciou o serviço de correlação é responsável por esta requisição.

No passo 3 a mesma aplicação é responsável por requisitar ao servidor de consulta web as frequências individuais de todos os termos, através de uma requisição via objeto JSON. A resposta do servidor de indexação é o número de documentos em que o termo se encontra. A requisição individual por ano ao servidor web é representada na Figura 18.

Figura 18 – Objeto *json*.



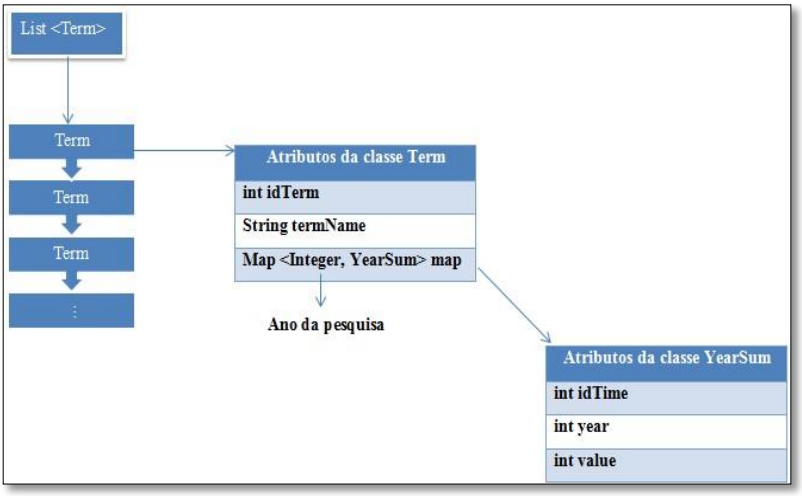
Fonte: Autor.

A Figura 18 representa a estrutura do objeto *json* enviado ao servidor web. A estrutura de consulta é composta pelo campo *project* que representa os campos em que a pesquisa será realizada no documento. O campo *operation* determina que o mesmo é um objeto de consulta. O campo *query* representa a consulta em si, ou seja, é passado o termo e em quais campos do documento indexado será realizada a pesquisa. O campo *year* é responsável pelo ano em que a pesquisa será realizada e o campo *offset* representa o deslocamento, cujo tamanho é 10. A requisição JSON indica mais informações no retorno, contudo, para o presente trabalho somente a quantidade de documentos que mencionam o termo é relevante.

O valor de retorno é então atualizado no campo frequência da tabela FT_CONCEPT_TIME (Figura 14).

No passo 4, após todos os termos obterem a frequência individual, o nó principal irá realizar uma pesquisa na base de dados, afim de montar a estrutura que proverá todos os dados para que cada nodo da *grid* possa executar o cálculo de correlação. Visto que a base foi desenvolvida considerando a temporalidade entre os termos, a seguinte estrutura foi desenvolvida:

Figura 19 – Estrutura do processo de correlação.



Fonte: Autor.

A Figura 19 demonstra a estrutura desenvolvida para dar suporte ao modelo temporal. Primeiramente é instanciado um objeto *List*, sendo que cada posição do mesmo terá um objeto *Term*. O objeto *Term* é responsável por guardar a identificação do termo, o nome do termo e uma instância de *hashMap* cuja chave é o ano em que o termo é mencionado e um objeto *YearSum*, responsável por armazenar a identificação do ano na base de dados, o ano e o valor, sendo este valor a frequência individual coletada e armazenada no passo 3, correspondente a um período de tempo específico.

No passo 5 cada nodo da *grid* é responsável por requisitar ao servidor de consulta a frequência conjunta dos termos, ou seja, a quantidade de documentos em que os dois termos são mencionados conjuntamente. O processo de geração da frequência conjunta é

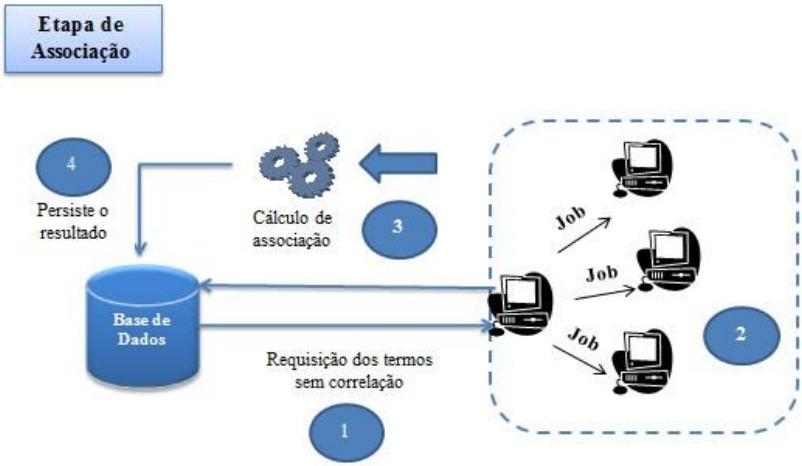
semelhante ao da frequência individual, porém é necessário enviar dois termos ao servidor de consulta. O resultado obtido é o número de documentos em que os dois termos ocorrem conjuntamente. Cada *Job* possui um termo origem (*source*) e uma lista de termos destino (*target*). Sendo assim, o *Job* calcula a frequência conjunta do termo origem com cada termo destino que compõe a lista. Com o valor da frequência individual e conjunta preenchidos é possível calcular o coeficiente de correlação. No passo 6 é efetuado o cálculo, através da fórmula *phi-squared*, apresentada no capítulo 2, tópico referente aos modelos baseados em coocorrência do presente trabalho.

Após obter o coeficiente de correlação cada *Job* é responsável também por persistir está informação na base de dados, passo 7. A quantidade de *Job*s gerados é igual ao número de termos presentes no objeto *List* menos um.

4.2.7 Detalhamento do processo de associação

A seguir o serviço de associação será descrito de maneira detalhada a partir da Figura 20 que representa a outra parte da arquitetura física.

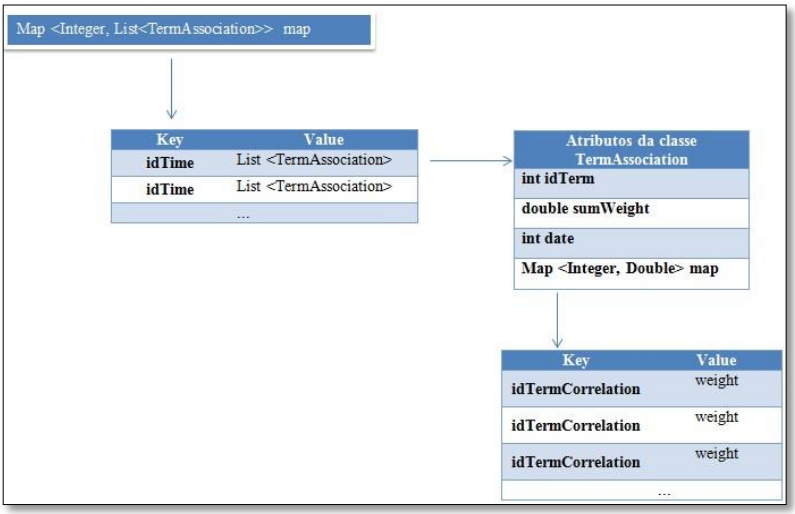
Figura 20 – Processo responsável pela associação.



Fonte: Autor.

Requisição dos termos sem correlação: O processo de descoberta de associação existente entre termos é realizado somente entre termos que não possuem correlação direta. No passo 1 verifica-se a requisição através de uma consulta ao banco de dados dos termos com seus respectivos termos relacionados, pela aplicação que iniciou o processo de associação. O banco de dados já possui todos os valores referentes ao processo de correlação armazenados. Para suportar o modelo temporal proposto a seguinte estrutura, representada pela Figura 21, foi desenvolvida:

Figura 21 – Estrutura do processo de associação.



Fonte: Autor.

De acordo com o intervalo de anos inseridos na base de dados, são recuperados todos os termos mencionados (*sources*) naquele ano específico, mais os termos relacionados ao *source*, ou seja, o vetor de contexto do *source*. A identificação do ano será a chave da *HashMap* que conterá em seu valor um objeto *List* sendo em que cada posição do mesmo conterá um objeto *TermAssociation*. A classe *TermAssociation* é composta pelos seguintes atributos:

int id: representará o *source*, ou seja, a identificação do termo (entendido como um documento) em questão, ao qual se deseja descobrir se existe associação com outro termo.

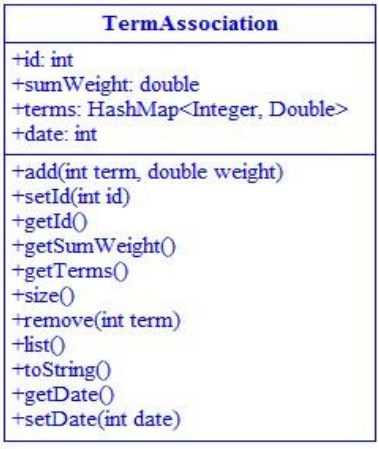
Map terms: representado através de uma *HashMap*, sendo que a parte referente a chave é um inteiro e a parte referente ao valor representa a correlação entre o termo em questão e o termo *source*. Desse modo, todos os termos inclusos na tabela *Hash* formam o vetor de contexto do termo *source*.

Double sumWeight: guarda a soma dos valores de correlação dos termos inclusos no vetor de contexto visando simplificar o processo do cálculo de associação.

int date: responsável por guardar o ano em que os termos estão associados.

Para um maior entendimento da classe *TermAssociation* a Figura 22 é apresentada. Nesta classe, o método *add* é responsável por carregar os termos visando formar o vetor de contexto. A *HashMap terms* recebe os temos e os valores de correlação de cada termo contido no documento. Ao atributo *sumWeight* é atribuído o valor de correlação elevado a potência de dois, ou seja, a cada nova entrada de termos é somado ao atributo *SumWeight* o valor de correlação contido no banco de dados elevado à potência de dois.

Figura 22 – Diagrama de classe *TermAssociation*.



Fonte: Autor.

Divisão das tarefas: O processo de associação consiste em verificar possíveis relacionamentos indiretos entre os termos. Após o término da operação de recuperação, o processo mestre envia à cada

nodo da *grid* uma posição da *map* já carregada, sendo esta posição representada por um ano em específico, pois a chave da *map* é composta pelos anos em que a pesquisa é realizada.

Cálculo do coeficiente de associação: no momento da execução do *job* é realizada a instanciação de uma *HashMap* que terá como chave um objeto do tipo *TermAssociation* e como valor outro objeto do tipo *TermAssociation.* Antes da adição dos dados na tabela *Hash* auxiliar, é realizado um teste para verificar a existência de correlação entre os termos em questão. A descoberta é feita perguntando-se a chave da *Hash* principal se a identificação de determinado documento consta como chave. Caso não exista é então realizada a adição na *Hash* auxiliar da posição em específico com a posição seguinte da *Hash.* Com a *Hash* auxiliar devidamente carregada é efetuado o cálculo, passo 3 da Figura 20. Aplicar-se-á o cálculo apresentado no capítulo 2, tópico referente ao modelo vetorial.

Armazenamento dos valores: No passo 4 cada nodo é responsável por fazer a inserção do resultado do valor do cálculo de associação na base de dados. O campo para armazenamento será (ASSOCIATION_STRENGTH), da tabela FT_RELATION_TIME, conforme apresentado na Figura 14.

A fim de uma maior explanação o seguinte exemplo, apresentado na Tabela 6 foi elaborado e testado utilizando o protótipo desenvolvido e o banco de dados *PostgreSQL* para armazenamento dos dados.

Tabela 6-Matriz de teste considerando o protótipo desenvolvido.

Documento					
	Biotecnologia	Engenharia genética	Tecnologia	Transgênico	Genética
Biotecnologia	0.0	0.7	0.9	0.0	0.0
Engenharia genética	0.7	0.0	0.3	0.6	0.8
Tecnologia	0.9	0.3	0.0	0.0	0.0
Transgênico	0.0	0.6	0.0	0.0	0.66
Genética	0.0	0.8	0.0	0.66	0.0

Fonte: Autor.

A matriz é montada considerando uma situação hipotética, utilizando-se nomes de pessoas e o peso de correlação existente entre os mesmos. Com base nesta matriz, armazenada no banco de dados, é feito uma seleção, via o protótipo, para recuperar todos os ids (identificação do termo no banco) dos documentos de maneira distinta. Esta seleção é realizada pelo nodo da grid que iniciou a aplicação. Após o término da mesma é iniciada a divisão dos *Jobs*, ou seja, são repassados a cada nodo pertencente à *grid* os vetores de contextos ao qual deverão aplicar o cálculo. Os vetores de contexto extraídos da matriz acima serão os seguintes:

- Biotecnologia {Engenharia genética, 0.7; Tecnologia, 0.9; Transgênico, 0.0; Genética, 0.0};
- Engenharia genética {Biotecnologia, 0.7; Tecnologia, 0.3; Transgênico, 0.6; Genética, 0.8};
- Tecnologia {Biotecnologia, 0.9; Engenharia genética, 0.3; Transgênico, 0.0; Genética, 0.0};

- Transgênico {Biotecnologia, 0.0; Engenharia genética, 0.6; Tecnologia, 0.0; Genética, 0.66};
- Genética {Biotecnologia, 0.0; Engenharia genética, 0.8; Tecnologia, 0.0; Transgênico, 0.66};

Estes vetores serão adicionados a uma lista. No momento da divisão dos *Jobs* o primeiro nodo da grid calculará o primeiro vetor com os demais contidos na lista, o segundo nodo calculará o segundo vetor com os demais da lista, o terceiro nodo calculará o terceiro vetor com os demais e assim por diante. No momento da realização do cálculo de associação o algoritmo executará outra função importante: verificar se determinado termo em questão não está relacionado com o termo que se deseja realizar o cálculo, visto que, para ocorrer a associação é necessário que os termos nunca tenham sido mencionados conjuntamente em um mesmo documento.

Com base nos vetores gerados anteriormente pode-se observar os termos que nunca foram mencionados em conjunto, ou seja, onde o peso de correlação é 0.0. Com isso o algoritmo calculará o vetor do termo Biotecnologia com o vetor do termo Transgênico e com o vetor do termo Genética. Será calculado também o vetor do termo Tecnologia com o vetor do termo Transgênico e com o vetor do termo Genética.

Com a aplicação do cálculo apresentado no capítulo 2, tópico referente ao modelo vetorial obteve-se o seguinte resultado:

Tabela 7-Resultado da aplicação do cálculo de associação pelo protótipo.

Termo	**Termo**	**Peso de Associação**
Biotecnologia	Transgênico	0.412
Biotecnologia	Genética	0.473
Tecnologia	Transgênico	0.212
Tecnologia	Genética	0.243

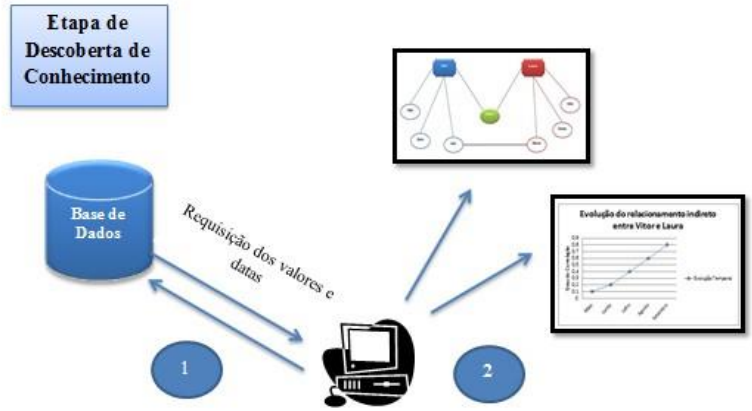
Fonte: Autor.

Na primeira coluna da matriz observa-se o termo origem (*source*), na segunda coluna o termo destino (*target*) e na terceira coluna o peso de associação existente entre o termo origem e o destino. Com isso obtém-se somente o peso de associação entre os termos que nunca foram mencionados conjuntamente em um mesmo documento, porém devido ao contexto no qual estão inseridos configuram um grau de associação por possuírem outros termos em comum.

4.2.8 Detalhamento do processo de descoberta de conhecimento

Após os etapas de correlação e de associação serem concluídas, os dados encontram-se devidamente armazenados no banco de dados. A partir disto é possível realizar estudos visando extrair informações que possibilitem, por exemplo, a análise de cenários que envolvam a descoberta de tendências. Em razão de a arquitetura permitir o armazenamento da data em que o processo foi executado, pode-se realizar um acompanhamento da possível evolução ou não do grau de associação entre dois termos quaisquer. No processo de descoberta de conhecimento uma aplicação acessa a base de dados onde se encontram os valores obtidos com a aplicação da etapa 1 da Figura 23. Através destes valores e da data dos mesmos pode-se, por exemplo, gerar mapas de tópicos, gráficos (etapa 2) conforme apresentado no Capítulo 2, tópico referente a associação de elementos textuais.

Figura 23 – Etapa de descoberta de conhecimento.



Fonte: Autor.

5 APRESENTAÇÃO DOS RESULTADOS

Neste capítulo será realizada a apresentação dos resultados, objetivando promover a explanação dos processos de correlação e associação e permitir a avaliação e discussão dos resultados obtidos com a aplicação do protótipo. A base de dados foi construída com base em artigos coletados na revista *ScienceDirect*[12].

5.1 INTRODUÇÃO

As discussões envolvendo este capítulo estão divididas em três partes, sendo:

Cenário da Aplicação: nesta parte é apresentado de maneira geral o cenário envolvendo as características da coleta de dados. Declara a abrangência do cenário sobre a base de dados, ou seja, quais tabelas estão envolvidas no processo. Promove uma visão geral das possibilidades de análise a partir do modelo e do cenário de aplicação.

Análise de Perfil: nesta segunda parte é realizada uma introdução sobre a análise envolvendo a correlação e a associação, e o seu resultado visual através de histogramas e gráficos. Apresenta casos de análise e expõe a análise entre termos em um contexto temporal.

Mapa de Tópicos: na terceira parte é provida uma visão sobre a análise de mapa de tópicos e sua importância como uma ferramenta para entender determinado contexto/domínio de aplicação. Neste momento, serão explanados os resultados para verificar a validade e consistência do protótipo desenvolvido e avaliar se o mesmo atingiu seus objetivos.

5.2 CENÁRIO DE APLICAÇÃO

O cenário de aplicação envolveu a coleta de artigos na base de dados da revista *ScienceDirect.* Os fatores que influenciaram na decisão da base de dados foram: a abrangência de áreas; por ser uma base de periódicos relevante no cenário de pesquisas mundial; pelo volume de artigos publicados, aproximadamente 11.770.732 e devido aos filtros de pesquisa, pois através dos mesmos os resultados foram mais expressivos para os fins da coleta.

[12] http://www.sciencedirect.com/

A extração dos dados para formar a base compreendeu o dia 30 e 31 de maio e visando aumentar a base foram coletados mais artigos no dia 05 de junho. Como o objetivo deste trabalho é evidenciar a relação temporal existente entre termos de determinado domínio, onde os mesmos envolvem determinados contextos, optou-se por montá-la conforme a Tabela 8:

Tabela 8-Tabela de pesquisa para montagem da base de dados.

Termo de Pesquisa 1	**Termo de Pesquisa 2**	**Período de realização da pesquisa**	**Momento em que ocorre a correlação**
Biotechnology	Genetic Engineering	1993 a 2002	2003
Nanotechnology	Medicine	1984 a 1993	1994

Fonte: Autor.

Cabe salientar que o trabalho não possui o acesso completo a base de dados, ou seja, não foi possível extrair todos os artigos da base compreendidos neste período. Este trabalho procurou evidenciar o aumento na associação e posteriormente a correlação envolvendo os termos *Biotechnology* e *Genetic engineering*, conjuntamente e, por conseguinte os termos *Nanotechnolog* e *Medicine*. As coocorrências que ocorreram antes do momento de correlação apresentado na Tabela 8, coluna quatro, foram desconsideradas, pois as mesmas ocorrem com baixa frequência conjunta.

Os termos foram mantidos em inglês, devido ao fato da pesquisa dos termos ocorrerem em inglês. Na primeira coluna da Tabela 8 encontram-se o termo para a realização da pesquisa, assim como na segunda coluna. A terceira coluna corresponde ao período de coleta dos documentos e a quarta coluna relata o ano em que os termos de pesquisa, coluna um e coluna dois, passaram a ser mencionados conjuntamente, ou seja, o momento em que ocorreu a correlação dos termos. Os documentos compreendidos dentro do período de pesquisa foram selecionados e distribuídos ao longo do tempo, ao qual foram coletados 313 documentos para o primeiro estudo de caso e 239 documentos coletados para o segundo estudo de caso. A pesquisa foi

realizada considerando a presença do termo no documento como um todo.

Com base nas informações da Tabela 8 foi montada a base de dados. Para este processo foi necessário extrair os dados relevantes dos artigos e estruturá-los na forma de documentos XML[13]. A criação dos documentos XMLs compreendeu o período entre 30/05/2013 a 07/06/2013, devido ao fato da extração dos dados ter ocorrido manualmente, para o primeiro estudo de caso, e no período entre 06/06/2013 a 10/06/2013, para o segundo estudo de caso. A estrutura do XML é apresentada na Figura 24:

Figura 24 – Estrutura do XML.

```
<DOCUMENT
        ID = ""
        TITLE = ""
        YEAR = "" >

        <AUTHORS>
                <ITEM NAME = "" ORGANIZATION = ""/>
        </AUTHORS>

        <KEYWORDS>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
                <ITEM NAME = ""/>
        </KEYWORDS>
</DOCUMENT>
```

Fonte: Autor.

Como demonstra a Figura 24 foram extraídos dos documentos o identificador, ao qual foi criado um sequencial, o título, o ano, o nome do(s) autor(es) com sua(s) respectiva(s) organização(ões) e as palavras-chave. Caso as palavras-chave não existissem, o documento era lido e as palavras elencadas no arquivo XML correspondente.

Após a anotação manual dos documentos, passou-se para a segunda etapa, ou seja, extrair via aplicação os dados do XML e criar um objeto JSON com os mesmos. Após o JSON concluído é necessário enviá-lo a um servidor de indexação que com base no objeto e no texto

[13] Segundo a W3Schools (2013), o XML é uma linguagem de marcação e foi projetado para transportar dados, não para exibir dados.

extraído do documento que se encontra no formato PDF, irá realizar a indexação do mesmo.

Com a indexação finalizada a aplicação enviará os termos indexados adicionados em uma *HashSet*, objeto Java, ao método *insertBase()* da classe *CreateBase()*, da aplicação, para que o mesmo possa realizar o processo de inserção dos dados na base de dados *Postgres*. Os dados serão inseridos na tabela DI_TERM do banco de dados, conforme a Figura 25.

Figura 25 – Inserção na tabela DI_TERM.

	term_id integer	description character varying(255)	normalized_description character varying(255)
1	1	Host resistance	
2	2	Abiotic stress	
3	3	Transgenic tobacco	
4	4	Plastids	
5	5	Immunoglobulin gene	
6	6	Molecules	
7	7	GM technology	
8	8	Microbial	
9	9	Exposure assessment	
10	10	Toxic metals	

Fonte: Autor.

Ao total foram 710 termos inseridos na base de dados para o primeiro caso e 506 termos para o segundo caso. O critério de seleção dos termos foi com base nas palavras chaves dos artigos coletados na base de dados. Objetivando a variação temporal dos artigos, foi inserido um total de 21 anos na tabela DI_TIME, tanto para o primeiro, quanto para o segundo caso conforme a Figura 26.

Figura 26 – Inserção na tabela DI_TIME.

	time_id integer	day integer	month integer	year integer
1	0	-1	-1	-1
2	1	0	0	1993
3	2	0	0	1994
4	3	0	0	1995
5	4	0	0	1996
6	5	0	0	1997
7	6	0	0	1998
8	7	0	0	1999
9	8	0	0	2000
10	9	0	0	2001

Fonte: Autor.

A construção de um cenário que visa à variação temporal é justificada pelo fato da análise da correlação e associação serem realizada em médio e longo prazo. O ano -1 foi criado a fim de armazenar a frequência total do termo na base, ou seja, a soma de todos os anos.

As classes e os domínios dos termos foram definidos como genéricos devido ao mecanismo de busca utilizado (servidor de consulta) não possuir semântica, tornando assim desnecessário o contexto das palavras. Caso o serviço de consulta possuísse semântica as classes e domínios dos termos seriam definidos de acordo com o contexto dos mesmos.

Na tabela FT_CONCEPT_TIME são armazenadas as frequências individuais dos termos. A frequência de cada termo é fundamental para o processo de correlação. A Figura 27 demonstra o conteúdo desta tabela.

Figura 27 – Inserção na tabela FT_CONCEPT_TIME.

	concept_id integer	time_id integer	frequency double precision
1	1	0	4
2	1	6	1
3	1	9	1
4	1	13	1
5	1	14	1
6	2	0	51
7	2	7	1
8	2	8	1
9	2	9	2
10	2	10	5

Fonte: Autor.

A tabela FT_RELATION_TIME é atualizada frequentemente durante a execução do processo de correlação e de associação. Ao realizar o processo de correlação foram geradas 1.552.107 entradas/tuplas na tabela, sendo deste total, 921.284 para a correlação e 630.823 para a associação, no primeiro caso. No segundo caso foram geradas 509.822 entradas/tuplas, sendo este valor composto por 302.421entradas/tuplas para a correlação e 207.401entradas/tuplas para a associação. A Figura 28 representa esta tabela.

Figura 28 – Inserção na tabela FT_RELATION_TIME.

	source_id integer	target_id integer	relation_id integer	time_id integer	correlation_strength numeric	association_strength numeric	joint_frequency integer
1	94	97	1	11	0.099999999974	0	1
2	98	103	1	12	0.166666666643333	0	1
3	96	97	1	5	0.083333333305	0	1
4	80	81	1	17	0.0833333333306666	0	1
5	98	97	1	12	0.0666666666637333	0	1
6	1	3	1	13	0.0624999999725	0	1
7	83	84	1	14	0.0454545454263 63	0	1
8	4	3	1	10	0.0833333333306666	0	1
9	93	99	1	8	0.24999999998	0	1
10	2	3	1	10	0.081632653004081	0	2

Fonte: Autor.

Esta tabela possibilita a persistência da frequência conjunta e do coeficiente de correlação, assim como permite armazenar o grau de associação entre os termos. As tabelas apresentadas são utilizadas como suporte ao processo de análise. A seguir serão descritos alguns tipos de análise e representações gráficas que podem ser produzidas a partir dos resultados oriundos dos serviços de correlação e associação.

Para a execução do cenário foi utilizado um computador pessoal, com processador Intel core i5 segunda geração, HD de 640 GB, 6 GB de memória e *clock* do processador de 2,40 GHz.

5.3 EXPLANAÇÃO DOS RESULTADOS

O modelo proposto suporta a evolução temporal existente entre termos que representam determinado domínio de análise. Tanto o processo de correlação quanto o processo de associação, levam em consideração o fator temporal.

Ao realizar uma análise de redes sobre determinado contexto o entendimento do mesmo é facilitado. A análise elenca um termo origem e gera um mapa com os termos mais significativos relacionados ao mesmo. Pode ser analisado ainda o contexto que circunda o termo relacionado a uma determinada origem, expandindo assim o mapa de tópicos que será gerado.

Como resultado deste mapeamento obtém-se gráficos que facilitam a visualização e entendimento do contexto dos termos. A

seguir serão apresentados alguns gráficos e mapas de tópicos referentes à aplicação do modelo proposto.

5.3.1 Análise de perfil

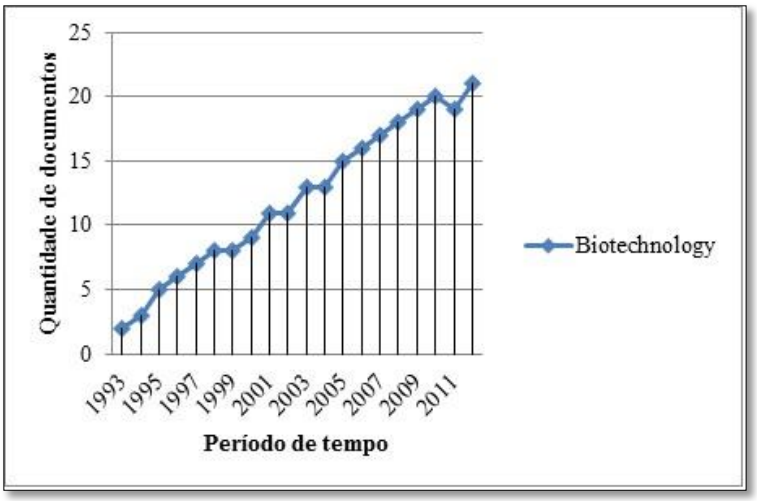
A seguir serão apresentadas as análises de perfil, representadas por gráficos, assim como, uma discussão dos mesmos.

5.3.1.1 Análise da frequência individual dos termos

A frequência individual dos termos representa o número de documentos que mencionam o termo em específico na base de dados. Abaixo serão apresentados os gráficos referentes aos termos descritos na Tabela 8, considerando a relação tempo.

No Gráfico 2 e no Gráfico 3 pode-se observar um aumento na quantidade de vezes em que os termos *Biotechnology* e *Genetic engineering* aparece nos documentos coletados.

Gráfico 2 – Frequência individual do termo *Biotechnology.*



Fonte: Autor.

Gráfico 3 – Frequência individual do termo *Genetic engineering*.

Fonte: Autor.

Abaixo serão apresentados o Gráfico 4 e o Gráfico 5 referentes aos termos Nanotechnology e Medicine.

Gráfico 4 – Frequência individual do termo *Nanotechnology*.

Fonte: Autor.

Gráfico 5 – Frequência individual do termo *Medicine.*

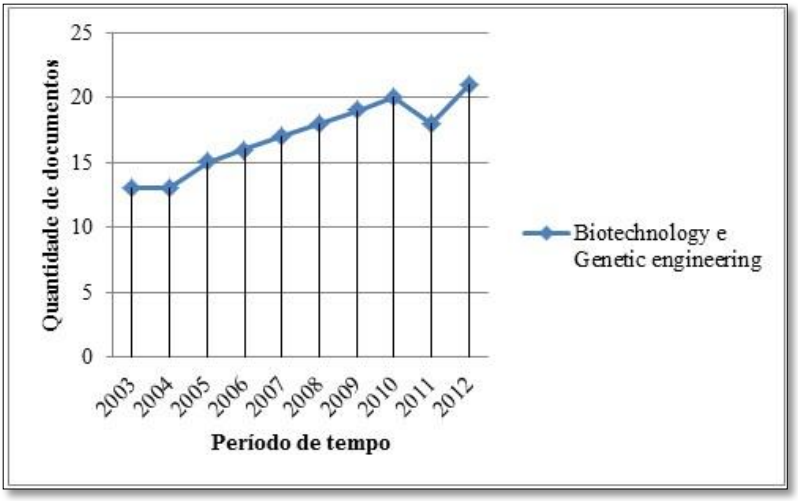
Fonte: Autor.

Tanto o Gráfico 4, quanto o Gráfico 5 demonstram um aumento na frequência individual dos termos pesquisados, apesar de algumas variações e decréscimos significativos em alguns anos.

5.3.1.2 Análise da frequência conjunta dos termos

A frequência conjunta representa a quantidade de documentos que dois termos aparecem conjuntamente. Com base nisto é possível acompanhar a evolução da frequência conjunta dos termos, como exemplifica o Gráfico 6 e o Gráfico 7.

Gráfico 6 – Frequência conjunta dos termos *Biotechnology* e *Genetic engineering.*



Fonte: Autor.

Gráfico 7 – Frequência conjunta dos termos *Nanotechnology* e *Medicine.*

Fonte: Autor.

O Gráfico 6 demonstra a evolução da frequência conjunta dos termos *Biotechnology* e *Genetic engineering* de maneira praticamente

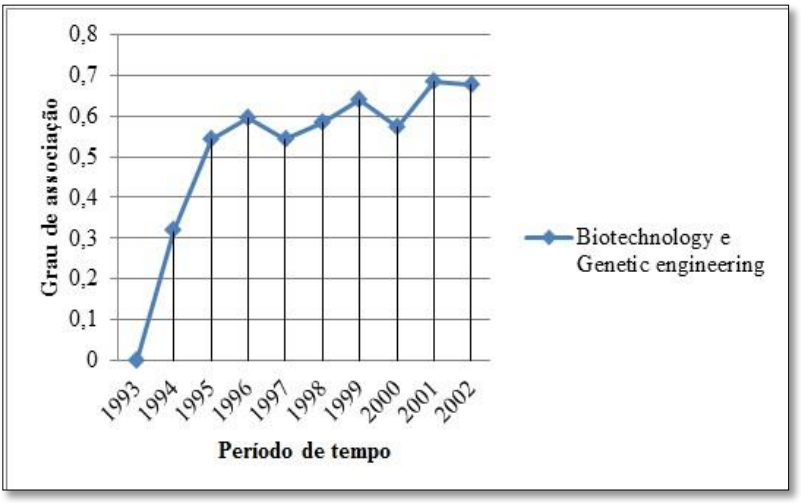
crescente entre os anos. Por outro lado, o Gráfico 7 demonstra algumas oscilações entre as frequência conjuntas de *Nanotechnology* e *Medicine*, mas tende a evolução.

5.3.1.3 Grau Associação

A associação é a proximidade entre dois termos, antes de estes passarem a serem mencionados conjuntamente. Apesar de existir uma tendência no aumento da associação ao longo do tempo e tal fato ser passível de investigação em cenários de análise, não existe a garantia de que a correlação entre dois termos irá ocorrer. Abaixo serão apresentados os gráficos referentes aos termos descritos na Tabela 8, considerando a relação tempo.

No Gráfico 8 pode-se observar um aumento no grau de associação existente entre *Biotechnology* e o termo *Genetic engineering*. Esta análise é de suma importância, pois através da mesma é possível analisar e identificar a aproximação do momento em que poderá ocorrer a correlação entre dois termos, ou seja, é possível inferir quando os termos em específico poderão vir a serem mencionados conjuntamente.

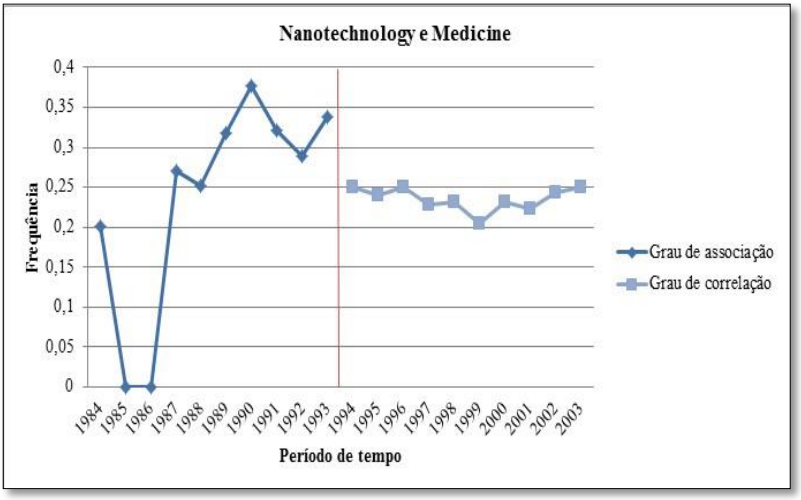
Gráfico 8 – Grau de associação entre *Biotechnology* e *Genetic engineering*.



Fonte: Autor.

O Gráfico 8 apresenta uma evolução associativa consistente tendendo a um (valor máximo). Entretanto, podem-se configurar casos em que a associação evolui de maneira discreta e, em um dado momento, os termos são correlacionados, ou seja, mencionados conjuntamente em um determinado documento. Significa que os termos mantinham certa associação e que em um dado momento, ainda que a associação seja baixa, a correlação ocorre. Este cenário pode ser vislumbrado no Gráfico 9 que representa os termos *Nanotechnology* e *Medicine*. Nos resultados, ainda que modesto, os valores de associação passam de 0,193 para 0,329, um incremento de aproximadamente 59%.

Gráfico 9 – Evolução temporal entre os termos *Nanotechnology* e *Medicine.*

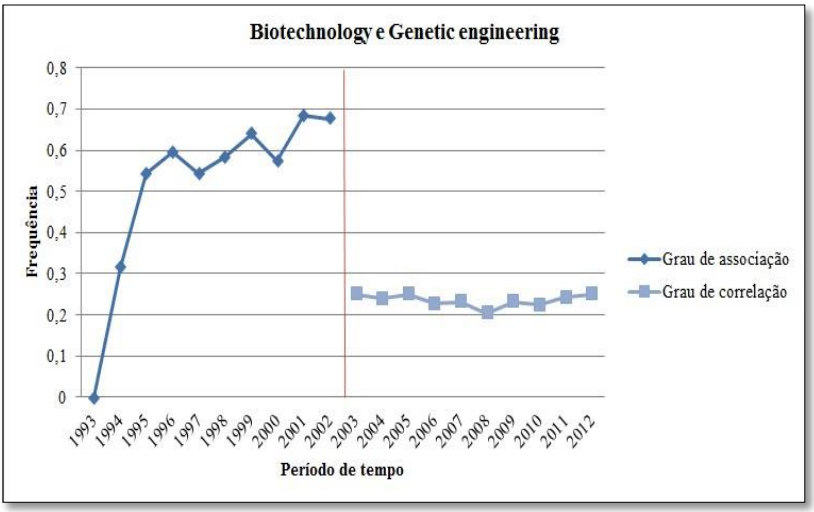


Fonte: Autor.

5.3.1.4 Grau de Correlação

O coeficiente de correlação mede o grau de correlação entre dois termos. Na equação utilizada no protótipo, *Phi-square*, apresentada no capítulo 2 do presente trabalho, são consideradas as frequências individuais e conjuntas descritas anteriormente. O Gráfico 10 é responsável por representar o momento de fusão entre os termos. Está sendo evidenciado o momento em que ocorre a associação entre os mesmo e o momento em que termina a associação e inicia-se a correlação.

Gráfico 10 – Evolução temporal entre *Biotechnology* e *Genetic engineering.*



Fonte: Autor.

5.3.2 Mapa de tópicos

Conforme explicitado, os termos mudam frequentemente o grau de associação e o grau de correlação entre os mesmo. Este fato se dá em virtude da relação temporal existente entre os mesmo e o contexto em que estão inseridos. Porém, existe a presença de termos no vetor de contexto de cada termo em destaque que proporcionam a associação entre os mesmos.

O vetor de contexto é responsável por descrever o termo. Cada termo possui o seu vetor de contexto, sendo que este contém todos os termos relacionados ao termo em destaque. Isto promove o contexto em que cada termo está inserido. Projetando-se os vetores dos termos de maneira gráfica, utilizando o conceito de redes, pode-se verificar visualmente a interconexões.

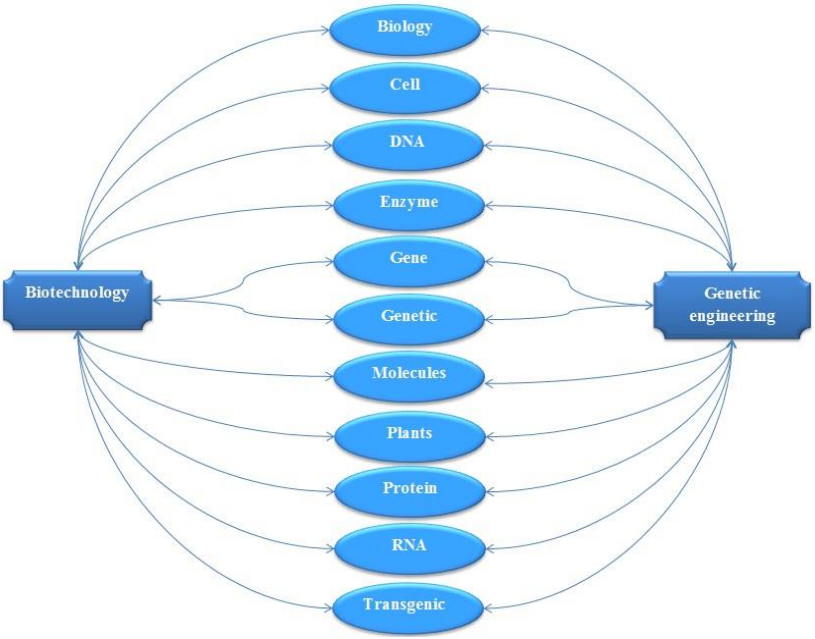
Buscando o entendimento do contexto de determinado termo de interesse, foi aplicada a projeção de redes, também chamada de mapa de tópicos. Esta análise elenca um termo origem e gera um mapa com os termos mais significativos do vetor de contexto do mesmo. A análise pode ainda ser expandida, ou seja, dado os termos mais significativos, em relação ao termo origem, estes também podem ser analisados.

Como resultado deste mapeamento obtém-se gráficos que facilitam a visualização e o entendimento do contexto dos termos.

A seguir será apresentado o mapa de tópicos obtido com base no relacionamento existente entre *Biotechnology* e *Genetic engineering*. O mapa gerado utiliza os conceitos mais significativos e que proporcionam a ligação entre documentos contendo os termos em destaque. Cabe salientar que o critério de seleção para a escolha dos conceitos, foi a presença dos mesmos, nos documentos coletados para o termo *Biotechnology* e para o termo *Genetic engineering*, sendo que os mesmos deveriam estar presente no momento da associação e no momento da correlação.

Abaixo será apresentada a Figura 29 que especifica o vetor de contexto compartilhado dos termos *Biotechnology* e *Genetic engineering*.

Figura 29 – Grafo de compartilhamento de características entre vetores de contexto.



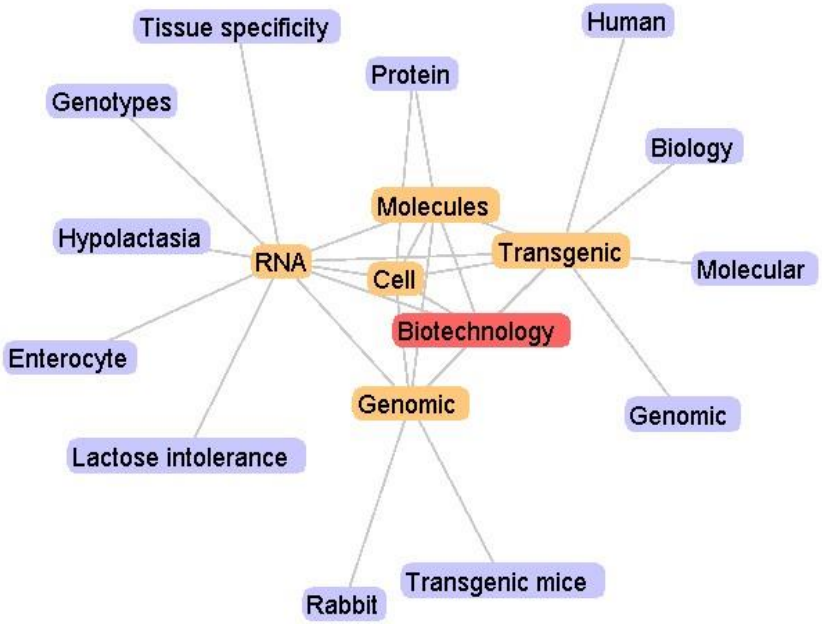
Fonte: Autor.

Na Figura 29 podem-se observar os termos que estão presentes no vetor de contexto do termo *Biotechnology* e no vetor de contexto do

termo *Genetic engineering*. Estes são os termos mais significativos e que proporcionam a conexão entre os documentos em questão.

A seguir serão apresentadas figuras com os mapas de tópicos envolvendo os termos em questão. Para a elaboração das mesmas foi utilizado como critério a escolha dos cinco termos mais correlacionados aos termos em questão, no caso, *Biotechnology* e *Genetic engineering*, e posteriormente foi realizada a seleção dos cinco termos mais correlacionados a estes, ou seja, uma expansão da rede considerando 2 níveis. A seguir será apresentada a Figura 30 com o mapa de tópicos gerado a partir do termo *Biotechnology*.

Figura 30 – Mapa de tópicos referente ao termo *Biotechnology*.



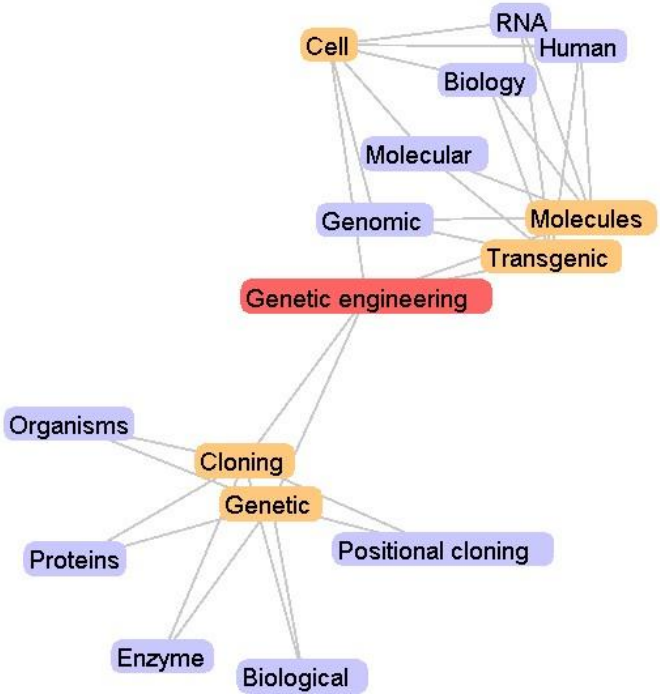
Fonte: Autor.

Como se pode observar os termos não possuem direção, ou seja, a correlação que existe entre dois termos quaisquer A, que se refere a *Biotechnology* e B, que se refere a *Genetic engineering*, e entre B e A, é a mesma. Outra característica importante é relação existente entre os termos contidos nos vetores de contextos do termo A com os termos do vetor de contexto de termo B. O que demonstra a proximidade dos termos dos documentos. Em destaque na cor vermelha encontra-se o

termo ao qual está sendo realizada a análise, e na cor amarela o primeiro nível correspondente ao termo central, no caso *Biotechnology*.

A Figura 31 apresenta o mapa de tópicos gerado a partir do termo *Genetic engineering*.

Figura 31 – Mapa de tópicos referente ao termo *Genetic engineering*.

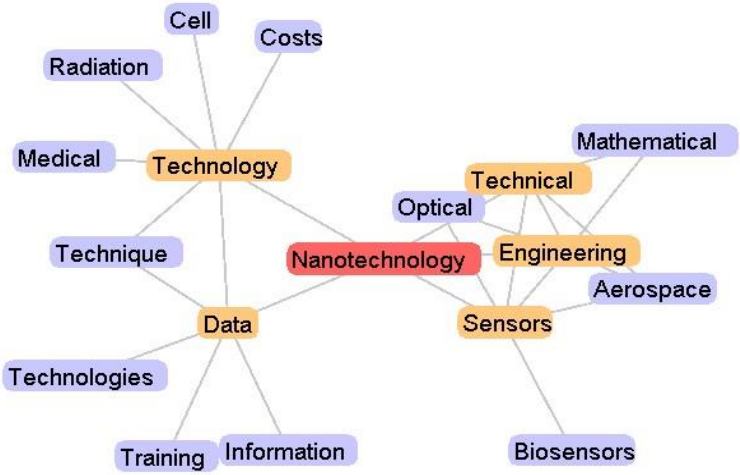


Fonte: Autor.

Os termos que promovem a associação entre os termos *Biotechnology* e *Genetic engineering* são: *Transgenic*, *Molecules* e *Cell*.

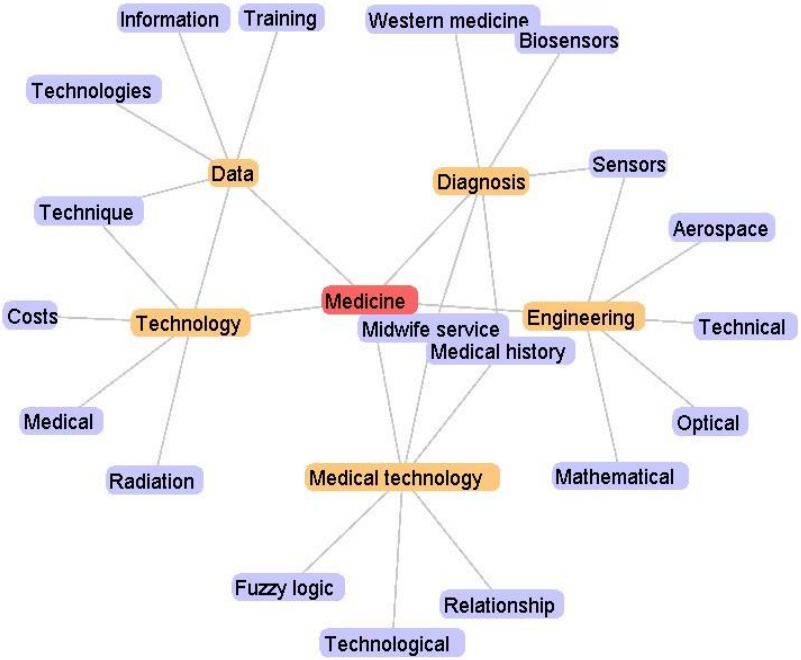
A Figura 32 e a Figura 33, representam os mapas de tópicos gerados a partir dos termos *Nanotechnology* e *Medicine*. Neste caso, os termos que promovem a associação são *Engineering*, *Technology* e *Data*.

Figura 32 – Mapa de tópicos referente ao termo *Nanotechnology*.



Fonte: Autor.

Figura 33 – Mapa de tópicos referente ao termo *Medicine*.



Fonte: Autor.

6 CONSIDERAÇÕES FINAIS

O objetivo geral do presente trabalho foi propor uma arquitetura computacional com base na computação distribuída e em técnicas de associação e correlação de elementos textuais que possibilitassem a descoberta de relacionamentos indiretos e temporais entre padrões textuais.

A partir da revisão das áreas de descoberta de conhecimento e uma revisão sobre a computação distribuída buscou-se entender o domínio e o contexto ao qual o trabalho aplicava-se. Com base nos conhecimentos adquiridos foi possível realizar estudos e desenvolver um protótipo que suportasse o objetivo geral deste trabalho.

O foco central do protótipo foi promover suporte aos processos de correlação e de associação considerando a sua aplicação em uma base de documentos que possibilitasse uma análise temporal. O processo de correlação utilizado é uma simplificação dos processos tradicionais de correlação, baseado no trabalho desenvolvido por Silva (2012), pois ao invés de inspecionar todos os documentos para verificar a quantidade de determinado padrão (termo) dentro destes documentos, considera-se somente a quantidade de documentos que mencionaram o termo. Quanto ao processo de associação o mesmo propõe-se em ser um modelo capaz de descobrir associações entre termos com base nos vetores de contextos dos mesmos. Este processo utiliza-se dos cálculos obtidos com a aplicação do processo de correlação.

Com base nos resultados foram geradas análises sobre o domínio desenvolvido, sendo as mesmas expostas através de gráficos temporais que evidenciavam a existência de padrões comportamentais entre os termos em análise. Adicionalmente, foram elaborados mapas de tópicos que vislumbrassem parte do vetor de contexto envolvendo o termo em análise. A arquitetura distribuída do protótipo demonstrou flexibilidade e escalabilidade podendo ser expandida quando necessário por meio de computadores com hardware e sistemas operacionais distintos.

Cabe salientar que o modelo para suportar os dados foi desenvolvido considerando o conceito de *Data Warehouse*. As tabelas que compõem o modelo são representadas por dimensões que fornecem o suporte para os dados e descrições. Têm-se ainda a presença de tabela de fatos, nestas tabelas são armazenadas medidas de valor que relacionem as tabelas de dimensões. Esse modelo pode em princípio representar qualquer domínio de aplicação que se baseie em relacionamentos entre conceitos. Entre as características centrais está a possibilidade de representação de relacionamentos em escala temporal.

As limitações encontradas durante o processo envolveram o cenário, pois foi considerado um conjunto restrito de tempos e documentos, visto que não era possível ter acesso à base por completo. Neste sentido, a base de dados foi construída manualmente. Outro ponto de limitação do trabalho se refere à execução do protótipo em que este não foi aplicado de maneira distribuída, ainda que forneça suporte para tal.

Durante o desenvolvimento do trabalho outras possibilidades foram vislumbradas como trabalhos futuros. Entre estas possibilidades menciona-se o desenvolvimento de uma interface gráfica que permita a integração com a aplicação e a base de dados, ao qual daria suporte à criação de gráficos e mapas de tópicos em geral para a apresentação das informações.

A base de conhecimento gerada pode tornar-se um apoio à colaboração de uma comunidade de práticas. Com isso a mesma caracteriza-se como um repositório de documentos, com base em um domínio, sendo possível fazer diversas buscas à base, por meio de sistemas computacionais. É possível ainda a utilização dos conceitos de análise de redes para uma interpretação mais apurada dos resultados gerados.

Pode-se vislumbrar ainda o desenvolvimento de um sistema para suportar a semântica existente entre os documentos, capaz de extrair padrões relevantes de determinado documento (chamados de entidades) e os relacionamentos entre estes. Apesar destes conceitos não terem sidos acoplados ao protótipo, o mesmo foi projetado pensando nestas futuras melhorias.

REFERÊNCIAS

AHMED, Kal; MOORE, Graham. Uma Introdução aos Mapas de Tópicos. **Architecture Journal**, 15 mar. 2006.

ARMBRUST Michael, et al. **A view of cloud computing**. *Commun. ACM*, v. 53, n. 4, p. 50-58, 2010.

BAKER, Mark; BUYYA, Rajkumar; HYDE, Dan. Cluster Computing: A High-Performance Contender. **IEEE Computer Society**, -, n. , p.79-83, jul. 1999.

BARÇANTE, Eduardo. **Propostas e metodologias de processamento automático de documentos textuais digitais: uma análise da literatura.** 2011. 101 f. Dissertação (Mestrado) - Universidade Federal Fluminense, Niterói, 2011.

BARION, Eliana Cristina Nogueira; LAGO, Decio. Mineração de Textos. **Revista de Ciências Exatas e Tecnologia**, Valinhos, Sp, p.123-140, 2008.

BARRETO, A. de A. **A eficiência técnica e econômica e a viabilidade de produtos e serviços de informação**. Rio de Janeiro, 1996.

BERMAN, F.; FOX, G. C.; HEY, A. J. G. **Grid Computing: Making the Global Infrastructure a Reality.** John Wiley& Sons, Inc, 2003.

BERNSTEIN, P. A. Middleware: **A Model for Distributed System Services**. Communications of the ACM, New York, v.3, n.2, p. 86-97, 1996.

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

BIEZUNSKI, Michel; NEWCOMB, Steve. **XML Topic Maps**: Finding aids for the web. 2001.

BOHN, Roger E.; SHORT, James E. **How Much Information?** Report on American Consumers. San Diego: -, 2009.

BOVO, Alessandro Botelho. **Um Modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais**. 155 p. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011

BRAGA, Ascenção. **Modernização do Sector Vitivinícola: Modelo de Sistema de Informação**. 1996. Dissertação (Mestrado) – Universidade da Beira Interior, Covilhã/Portugal, 1996.

BRANTNER, M., et al. Building a database on s3. In: **Proceedings of the 2008 ACM SIGMOD international conference on Management of data** - SIGMOD '08, page 251, New York. ACM Press. 2008.

BUYYA, R.; BROBERG, J.; BRANDIC, I. **Cloud computing and emerging it platforms**, 2009.

BUYYA, Rajkumar; VENUGOPAL, Srikumar. A Gentle Introduction to Grid Computing and Technologies. **Computer Society Of India, Mumbai**, p.09-19, 2005.

CECI, Flávio. **Um modelo semiautomático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados**. 2010. 131 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2010.

CHAVES, Marcirio Silveira. **Uma metodologia para a construção de ontologias e integração de conhecimento**. 2007. Tese (Doutorado) – Programa de Doutoramento em Informática da Universidade de Lisboa, Universidade de Lisboa, Portugal, 2007.

CHECKLAND, Peter; HOLWELL, Sue. **Information, Systems and Information Systems**, **Making Sense of the Field**, John Wiley & Sons, UK. 1998.

CHETTY, M; BUYYA, R. Weaving Computation Grids: How Analogous Are They with Electrical?. **Computing in Science and Engineering**, v.4, n.4, p. 61-71, 2002.

CHIAVENATO, Idalberto. **Introdução à teoria geral da administração.** 7ª edição Rio de Janeiro: Editora Campus, 2003.

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.

CONCEIÇÃO, Álvaro William da. **Um sistema voltado ao armazenamento e recuperação de conteúdo textual de diferentes contextos.** 2013. 61 f. Monografia (Graduação) - Curso de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, 2013.

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.

COULOURIS, G.; DOLLIMORE, J.; KINDBERG, T**. Distributed Systems: Concepts and Design**. 4ª ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

DANTAS, Mario A. R. **Computação distribuída de alto desempenho**: redes, clusters e grids computacionais. Rio de Janeiro: Axcel Books, 2005.

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

DAVENPORT, T. H. **Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação**. São Paulo: Futura, 1998.

DEITEL, H. M.; DEITEL, P. J.; CHOFFNES, D. R.. **Sistemas Operacionais**. 3ª edição São Paulo: Pearson Prentice Hall, 2005.

DEITEL, Harvey M. **Operating System**. 2º Edição, Framingham: Editora Bookman, 1990.

DEITEL, Harvey M.; DEITEL, Paul J. **Java como programar**. 8. ed. São Paulo (SP): Pearson Prentice Hall, 2010.

DOWNIE, N. M.; HEATH, R. W. **Basic statistical methods**. New York: Harper & Brothers, 1959.

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

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistema de Banco de Dados**. Revisor técnico Luíz Ricardo de Figueiredo. São Paulo: Pearson Addison Wesley. 2005

EPPLER, M. J. Making Knowledge Visible Through Intranet Knowledge Maps: Concepts, Elements, Cases. In: **Proceedings of the 34th Hawaii International Conference on System Sciences**, 2001

ETZIONI, O. Banko, M., Soderland, S. Weld D. S. **Open Information Extraction form the Web, Communications of the ACM**, 51(12):68-74, 2008.

FAYYAD, U. et al. **From Data Mining to Knowledge Discovery in Databases**. AAAI/MIT Press, 1997.

FELDMAN, R. et al. A domain independent environment for creating information extraction modules, In: **Proceedings of the tenth international conference on information and knowledge management**, p 581-588, ACM Press. 2001

FIALHO, F.A.P. et al. **Gestão do conhecimento e aprendizagem: as estratégias competitivas da sociedade pós-industrial**. Florianópolis: Visual Books, 2006.

FONSECA Filho, Cléuzio. **História da computação:** O Caminho do Pensamento e da Tecnologia. Porto Alegre : EDIPUCRS, 2007. 205 p.

FOSTER, Ian; KESSELMAN, Carl; TUECKE, Steven. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. **International Journal Of High Performance Computing Applications,** -, p. 200-222. 2001.

FOSTER, I. **The Grid: A New Infrastructure for 21st.** Century Science Physics Today, v. 55, p. 42-47, 2002.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J.. Knowledge Discovery in Databases: An Overview. **Al Magazine**, Palo Alto, v. 13, n. 3, p.57-70, 1992.

FREITAS, H. **As tendências em Sistemas de Informação com base em recentes congressos**. n.13. ReAd: Porto Alegre, 2000.

GANTZ, John; REINSEL, David. **The Digital Universe Decade – Are You Ready?** Framingham: Idc – Iview, 2010.

GONÇALVES, Alexandre L. ; UREN, Victoria ; KERN, Vinícius Medina ; PACHECO, Roberto C S . Mining Knowledge from Textual Databases: An Approach using Ontology-based Context Vectors. In: **International Conference on Artificial Intelligence and Applications (AIA 2005)**, 2005, Innsbruck. Proceedings of the 23rd IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, 2005. p. 66-71.

GONÇALVES, Alexandre Leopoldo**. Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento.** Florianópolis, SC, 2006. 196 f. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção.

GONÇALVES, Marcio. **Extração de dados para Data Warehouse**. 1ª ed. Rio de Janeiro: Axcel Book, 2003.

GOUVEIA, Roberta Macêdo Marques. **Mineração de dados em um data warehouse para sistema de abastecimento de água.** 2009. 147 f. Tese (Doutorado) - Universidade Federal da Paraíba, João Pessoa, 2009.

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

HAIR, J. F. et al. **Multivariate data analysis**. 5th. Prentice Hall; 5th edition, 1998.

HAN, J.; KAMBER, M. **Data Mining**: Concepts and techniques, Simon Fraser University, Morgan Kaufmann Publishers, 2000.

HEARST, Marti A.. Untangling text data mining. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37., 1999, Stroudsburg. **Proceedings... .** Stroudsburg: Association For Computational Linguistics, p. 3 – 10, 1999.

HILBERT, Martin; LÓPEZ, Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. **Science Magazine**, Califórnia, p. 60-65, 01 abr. 2011.

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.
IEEE Multimedia, p. 2-6, April-June 2001.

INMON, Willian H. **Builging the data warehouse**. 3. ed New York: J. Wiley, 2002. 412p.

IVANOV, Nikita; DMITRIY, Setrakyan. **Real Time Big Data Processing with GridGain.** Disponível em: <http://www.gridgain.com/book/book.html#_introduction>. Acesso em: 26 out. 2012.

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.

JÚNIOR, Elias Teodoro da Silva. **Middleware adaptativo para sistemas embarcados e de tempo real.** 2008. 127 f. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

JÚNIOR, Ivanilson França Vieira. **Uma abordagem na camada de middleware para troca dinâmica de componentes em sistemas multimídia distribuídos baseados no framework Cosmos.** 2009. 82 f. Tese (Doutorado) – Universidade Federal do Rio Grande do Norte, Natal, 2009.

JUNQUEIRA, Mirella Silva. **Uma proposta de interface de consulta para recuperação de informação em documentos semi-estruturados.** 2009. 125 f. Dissertação (Mestrado) - Universidade Federal de Uberlândia, Uberlândia, 2009.

KANTARDZIC, M. **Data Mining**: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003, 343 p.

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

KOCK, N.F. JR.; MCQUEEN, R.J.; SCOTT, J.L. (1997) Can Action Research be Made More Rigorous in a Positivist Sense? The Contribution of an Iterative Approach. **Journal of Systems and Information Technology**, v.1, n.1, p. 1- 24, 1997.

KORPEL, Eric J. et al. Status of the UC-Berkeley SETI Efforts. In: INSTRUMENTS, METHODS, AND MISSIONS FOR ASTROBIOLOGY, v. 14, 2011, San Diego. **Proceedings**... San Diego: Spie, 2011.

KOWALTOWSKI, Tomasz. Von Neumann: suas contribuições à Computação. **Estud. av**., São Paulo, v. 10, n. 26, Abr. 1996.

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.

LIMA, Carlos Eduardo de; FAGUNDES, Fabiano. Utilização de mapas de tópicos no desenvolvimento de hiperdocumentos educacionais. In: VI ENCONTRO DE ESTUDANTES DE INFORMÁTICA DO ESTADO DO TOCANTINS – ENCOINFO, 6., 2004, Palmas, Tocantins. **Anais**... Palmas, Tocantins: CEULP/ULBRA, 2004. 11 p.

LIRA, Sachiko Araki. **Análise de correlação**: abordagem teórica e de construção dos coeficientes com aplicações. 2004. 209 f. Dissertação (Mestrado) - UFPR, Curitiba, 2004.

LOUDON, Kyle. Desenvolvimento de grandes aplicações Web. São Paulo (SP): Novatec, 2010. 325 p.

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

LYMAN, Peter; VARIAN, Hal R. **How much information?** Executive summary. 2003.

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

MATTSSON, Michael. **Evolution and Composition of Object-Oriented Frameworks**. 2000. 231 f. Tese (Doutorado) - Departamento de Department Of Software Engineering And Computer Science, University Of Karlskrona/ronneby, Karlskrona, 2000.

MCGEE, J. V.; PRUSAK, L. **Gerenciamento estratégico da informação: aumente a competitividade e a eficiência de sua empresa utilizando a informação como uma ferramenta estratégica**. Rio de Janeiro: Campus, 1994.

MELLO, João Alexandre Bonin De. **UMA PROPOSTA DE MODELO DE DADOS PARA SUPORTE AO PROCESSAMENTO TRANSACIONAL E DE APOIO INFORMACIONAL SIMULTANEAMENTE.**2002. 101 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2002.

MOLLICK, E. Establishing Moore's Law. **Annals of the History of Computing**, v. 28, n. 3, p. 62 – 75, 2006.

MOONEY, Raymond J.; NAHM, Un Yong. Text Mining with Information Extraction. In: INTERNATIONAL MIDP COLLOQUIUM DAELEMANS, 4., September 2003, Bloemfontein, South Africa. W.,

du PLESSIS, T., SNYMAN, C. and TECK, L. (Eds.). **Proceedings**… Bloemfontein, South Africa: Van Schaik Pub., 2005. p.141-160.

MOORE, Gordon E. **Cramming more components onto integrated circuit**. Electronics Magazine. Ano. 8. N. 38, abr. 1965.

MORESI, E. A. D. **Delineando o valor do sistema de informação de uma organização**. Ciência da Informação, Brasília, v. 29, n. 1, 2000.

NURSEITOV, Nurzhan et al. Comparison of JSON and XML Data Interchange Formats: A Case Study. In: INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS IN INDUSTRY AND ENGINEERING, 22., 2009, Bozeman. **Proceedings**... San Francisco, California: Caine, 2009. p. 157 - 162.

O'LEARY, D. E. **Enterprise Knowledge Management**, IEEE Computer, v. 31 n.3 (March), pp.54-61, 1998.

OLIVEIRA, Alexandre; MARCOS, Adérito; VAASAN Thanigai. Topic maps na visualização de informação no ensino e treino. In: ACTAS DA PRIMEIRA CONFERÊNCIA DA ASSOCIAÇÃO PORTUGUESA DE SISTEMAS DE INFORMAÇÃO, 1., 2000, Minho, Portugal. **Proceedings**... Braga, Portugal: Universidade do Minho, out. 2000.

PACHECO, Roberto C. S. et al. Uma análise da pesquisa em engenharia e ciências mecânicas no Brasil a partir dos dados da Plataforma Lattes. **Associação Brasileira de Engenharia e Ciências Mecânicas (ABCM)**, v. 12, p.18-24, out. 2007.

PARK, J.; HUNTING, S. **XML topic maps:** creating end using topic maps for the web. Boston: Addison Wesley, 2003. 644 p.

RAJMAN, M.; BESANÇON, R. Text mining: Natural language techniques and text mining applications. **In: IFIP TC2/WG2.6 WORKING CONFERENCE ON DATABASE SEMANTICS (DS-7)**, 7., 1997, Leysin. Proceedings... .Chapman & Hall, 1997. p. 7 - 10.

RAMOS, Hélia de Sousa Chaves; BRÄSCHER, Marisa. **Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T. Ci. Inf.**, Brasília, v. 38, n. 2, p.56-68, ago. 2009.

RESENDE, Cristiano Marcelo. **Ambiente Grid utilizando Software Livre.** 2010. 62 f. Monografia (Especialização Lato Sensu) - Universidade Federal de Lavras, Lavras, 2010.

SANTOS, Rui Almeida. **Data Warehouse: Modelo de Auditoria e Controlo Interno.** 2009. 115 f. Dissertação (Mestrado) - Instituto Universitário de Lisboa, Lisboa, 2009.

SCHIESSL, José Marcello. **Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor.** 2007. 106 f. Dissertação (Mestrado) - Universidade de Brasília, Brasília, 2007.

SILBERSCHARTZ, Abraham; GALVIN, Peter Baer; GAGNE, Greg. **Fundamentos de Sistemas Operacionais.** 8ª edição Rio de Janeiro: LTC, 2010.

SILVA, E. R. G.; ROVER, A. J. O Processo de descoberta do conhecimento como suporte à análise criminal: minerando dados da Segurança Pública de Santa Catarina. In: **International Conference on Information Systems and Technology Management**, 2011, São Paulo. Anais da International Conference on Information Systems and Technology Management. São Paulo: FEA, 2011. v. 8.

SILVA, Guilherme Baião Salgado. **A UTILIZAÇÃO DE MAPAS DE TÓPICOS NA COMPATIBILIZAÇÃO DE CONTEÚDOS HIPERTEXTUAIS SEMANTICAMENTE ESTRUTURADOS.** 2008. 144 f. Dissertação (Mestrado) - Universidade Federal de Minhas Gerais, Belo Horizonte, 2008.

SILVA, Thales Nascimento da. **Uma Arquitetura para Descoberta de Conhecimento a partir de Bases Textuais.** 2012. 76 f. Monografia (Graduação) - Curso de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, 2012.

SILVA, Welmisson Jammesson da. **Extração de Informação Não Estruturada, Usando um Método Supervisionado e Independente de Domínio.** 2009. 92 f. Dissertação (Mestrado) - Universidade Federal de Campina Grande, Campina Grande, 2009.

SINGH, Harry S. Data Warehouse: Conceitos, Tecnologias, Implementação e Gerenciamento.1ª ed. São Paulo: Makron Books, 2001.

SMALHEISER, N. R. **Literature-Based Discovery: Beyond the ABCs**. Journal of the American Society for Information Science and Technology. doi: 10.1002/asi.21599. 2011.

SOROR, A. A., et al. **Automatic virtual machine configuration for database workloads**. ACM Trans. Database Syst., 35(1):1–47. 2010.

STALLINGS, William. **Arquitetura e organização de computadores**. 8. ed. São Paulo (SP): Pearson, 2010. xiv, 624p.

STEELS, L. Corporate knowledge management. **Proceedings** of ISMICK193, Compie`gne, France, pp. 9 – 30, 1993.

STEVENSON, William J. **Estatística aplicada a administração**. São Paulo: HARBRA, 2001.

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

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

TAN, A.-H. Text mining: The state of the art and the challenges. **In: Proceedings of the Pacifi c Asia Conference on Knowledge Discovery and Data Mining – PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases**, Beijing, p. 65–70, 1999.

TANENBAUM, Andrew S. **Sistemas operacionais modernos**. 3. ed. Rio de Janeiro (RJ): Prentice-Hall do Brasil, 2010.

TANENBAUM, Andrew S. **Sistemas Operacionais Modernos**. 3º Edição, São Paulo: Editora Campus,1992.

TANENBAUM, Andrew S.; STEEN, Maarten van. **Sistemas distribuídos: princípios e paradigma**. 2. ed. São Paulo: Pearson Prentice Hall, 2007.

TARDELLI, A. O.; ANÇÃO, M. S.; PACKER, A. L.; SIGULEM, D. **Descoberta baseada em literatura: um enfoque experimental para descoberta aberta em bases de dados do tipo MEDLINE**. In: CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE – CBIS, 8., 2002.

TAURION, Cezar. **Cloud Computing:** Computação em Nuvem: Transformando o mundo da tecnologia da informação.

TRYBULA, W. J. **Text mining. Annual Review of Information Science and Technology**, vol. 34, 1999, p. 385-419.

TUOMI, I. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organization memory. **Journal of Management Information Systems**, v. 16, n. 3, p. 103-117, 1999.

VIANNA, Rossana Cristina Xavier Ferreira et al . Mineração de dados e características da mortalidade infantil. **Cad. Saúde Pública**, Rio de Janeiro, v. 26, n. 3, Mar. 2010.

WEISS, S. M., et al. **Text Mining**: Predictive Methods for Analyzing Unstructured Information. Spring, New York, 2005.

WEISS, A. **Computing in the Clouds**. Networker, v.11, n. 4, p. 61-71, 2007.

WHITE, Tom. **Hadoop: The Definitive guide**. O'Reilly. p. 501, 2009.

WILKS, Yorick; CATIZONE, Roberta. Can We Make Information Extraction More Adaptive? In: SCIE99 WORKSHOP, 1999, Sheffield. **Proceedings**. Berlin: Springer-verlag, p. 1 – 16, 1999.

WIVES, Leandro Krug. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Orientador: Oliveira, José Palazzo Moreira de, 2004. 126f : il. Tese (Doutorado)-Universidade Federal do

Rio Grande do Sul. Porto Alegre: Programa de Pós-Graduação em Computação.

W3SCHOOLS. **Introduction to XML:** What is XML?. Disponível em: <http://www.w3schools.com/xml/xml_whatis.asp>. Acesso em: 25 mar. 2013.

ZORRINHO, C. **Gestão da Informação**. Condição para Vencer. Lisboa: Iapmei:1995.