

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO E SISTEMAS

WAGNER IGARASHI

**CONSTRUÇÃO AUTOMÁTICA DE VOCABULÁRIOS TEMÁTICOS E CÁLCULO
DE ADERÊNCIA CURRICULAR: UMA APLICAÇÃO AOS FUNDOS SETORIAIS**

Florianópolis, 2005

UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS

**CONSTRUÇÃO AUTOMÁTICA DE VOCABULÁRIOS TEMÁTICOS E CÁLCULO
DE ADERÊNCIA CURRICULAR: UMA APLICAÇÃO AOS FUNDOS SETORIAIS**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Universidade Federal de Santa Catarina como requisito parcial para obtenção do título de Mestre em Engenharia de Produção.

WAGNER IGARASHI

Orientador: José Leomar Todesco, Dr.

Florianópolis, 2005.

FICHA CATALOGRÁFICA

AUTOR:
IGARASHI, Wagner

Título: Construção automática de vocabulários temáticos e cálculo de aderência curricular: uma aplicação aos fundos setoriais

Número de folhas: 95

Dissertação de mestrado
Pós-graduação em Engenharia de Produção
Área de Concentração: Inteligência Aplicada

Orientador: José Leomar Todesco, Dr.

Palavras chave: vocabulários controlados; vetores de contexto; recuperação de informação.

Wagner Igarashi

**CONSTRUÇÃO AUTOMÁTICA DE VOCABULÁRIOS TEMÁTICOS E CÁLCULO
DE ADERÊNCIA CURRICULAR: UMA APLICAÇÃO AOS FUNDOS SETORIAIS**

**Esta dissertação foi julgada e aprovada para a obtenção do título de Mestre em
Engenharia de Produção no Programa de Pós-Graduação em Engenharia de
Produção e Sistemas da Universidade Federal de Santa Catarina.**

Florianópolis, 28 de Março de 2005.

**Prof. Edson Pacheco Paladini, Dr.
COORDENADOR DO CURSO**

Banca Examinadora:

**Prof. José Leomar Todesco, Dr.
Orientador**

Prof. Roberto C. S. Pacheco, Dr.

Prof. Aran Bey Tcholakian Morales, Dr.

Alexandre Leopoldo Gonçalves, Me.

*"Não me desencorajo, porque a cada tentativa errada descartada é outro passo à frente".
(Thomas Edson)*

"Ou nós encontramos um caminho, ou abrimos um". (Aníbal)

Por estar ao meu lado, em todos os momentos, contínua e incansavelmente me apoiando com seu amor e carinho, dedico este trabalho à minha esposa Deisy.

AGRADECIMENTOS

Para desenvolver a dissertação, precisei de muito empenho bem como do apoio e compreensão das pessoas que de forma direta ou indireta participaram de todo o seu processo de elaboração. Assim, manifesto meus sinceros agradecimentos:

a minha família pelo amor, carinho, atenção dispensados durante toda minha vida;

ao Prof. José Leomar Todesco, Dr., Prof. Roberto Pacheco, Dr., Prof. Aran Bey Tcholakian Morales, Dr. e Prof. Vinícius Medina Kern, Dr., pela oportunidade, confiança e atenção despendidos;

a Alexandre Leopoldo Gonçalves, Me. pela co-orientação, apoio e amizade;

a Marcos Luiz Marchezan, Me. pelo apoio e amizade;

a Denílson Sell, Me. pelo apoio e amizade;

à Universidade Federal de Santa Catarina, especialmente ao Departamento de Engenharia de Produção e Sistemas, pela oportunidade de realização do mestrado;

aos professores do Programa de Pós-Graduação em Engenharia de Produção e Sistemas, pelos conhecimentos transmitidos, que contribuíram para o desenvolvimento deste trabalho;

ao Grupo STELA por contribuir e disponibilizar as informações necessárias para o desenvolvimento deste trabalho;

aos professores e amigos do Curso de Bacharel em Informática da UEM – Universidade Estadual de Maringá, especialmente à: Maria Madalena Dias, Dra., pela motivação, apoio e incentivo que sempre transmitiram;

aos demais amigos e a todas as pessoas que de alguma forma contribuíram para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS.....	9
LISTA DE SIGLAS.....	11
GLOSSÁRIO.....	13
RESUMO	16
ABSTRACT	17
CAPÍTULO I – INTRODUÇÃO.....	18
1.1 Contextualização	18
1.2 Problematização	21
1.3 Objetivos	22
1.3.1 Objetivo geral	22
1.3.2 Objetivos Específicos	22
1.4 Justificativa	23
1.5 Contextualização do trabalho na Engenharia de Produção.....	24
1.6 Metodologia.....	25
1.7 Estrutura do trabalho	29
CAPÍTULO II –BUSCA E RECUPERAÇÃO DE INFORMAÇÃO	30
2.1 Considerações iniciais.....	30
2.2 Extração, recuperação e armazenamento das informações.....	30
2.3 Modelo de recuperação de informação espaço-vetorial	33
2.4 Indexação e normalização	36
2.4.1 Identificação de termos	37
2.4.2 Remoção de palavras irrelevantes (stopwords)	38
2.4.3 Normalização morfológica	39
2.4.3.1 <i>Stemming</i>	40
2.4.3.2 <i>N</i> -gramas	42
2.4.4 Cálculo de relevância do termo	44
2.4.5 Redução de dimensionalidade	45

2.5 Busca e recuperação	46
2.5.1 Vocabulários.....	47
2.5.2 Tesouros	48
CAPÍTULO III - SISTEMAS DE INFORMAÇÃO	50
3.1 Considerações iniciais	50
3.2 Conceituação dos sistemas de informação	50
3.3 Técnicas para a construção de sistemas de apoio à decisão	55
3.3.1 <i>Data warehouse</i> (DW).....	57
3.3.2 <i>Knowledge Discovery in Database</i> (KDD).....	59
3.3.3 <i>Knowledge Discovery in Text</i> (KDT).....	61
3.4 Modelo Espiral de Desenvolvimento de Software	62
CAPÍTULO IV – GERAÇÃO DE VOCABULÁRIOS A PARTIR DA PLATAFORMA LATTES	65
4.1 Considerações iniciais	65
4.2 Fomento e planejamento em C&T	65
4.2.1 Fomento em C&T no Brasil	66
4.3 Plataforma Lattes	68
4.4 Fundos setoriais	69
4.5 Modelo de sistema de apoio à decisão	70
4.5.1 Vocabulários temáticos Lattes.....	72
4.5.2 Busca textual com cálculo de aderência	79
4.5 Resultados	83
CAPÍTULO V – CONCLUSÕES E RECOMENDAÇÕES	88
5.1 Conclusões	88
5.2 Recomendações	90
REFERÊNCIAS BIBLIOGRÁFICAS	92

LISTA DE FIGURAS

Figura 1: Número de instituições, grupos, pesquisadores e doutores.....	19
Figura 2: Método para desenvolvimento e implementação do estudo	26
Figura 3: Estrutura básica de um índice de arquivo invertido.....	32
Figura 4: Modelo espaço-vetorial	35
Figura 5: Inter-relação dos tipos de sistemas com os níveis organizacionais e as informações geradas.....	53
Figura 6: Distinção dos SAD e SIG em relação aos focos	54
Figura 7: Uma visão dos passos que compõem o processo KDD	60
Figura 8: Modelo espiral de desenvolvimento de software.....	63
Figura 9: Modelo do sistema de apoio à decisão	71
Figura 10: DW da Plataforma Lattes	73
Figura 11: Integração do DMVT ao DW da plataforma Lattes.....	74
Figura 12: Data mart vocabulários temáticos	75
Figura 13: Sistema gerador de vocabulários temáticos (SGVT)	75
Figura 14: Site dos vocabulários temáticos - Consulta.....	77
Figura 15: Site dos vocabulários temáticos - Resultado.....	77
Figura 16: Site de busca textual – Consulta.....	81
Figura 17: Site de busca textual – Resultado.....	82
Figura 18: Modelo preliminar da base dos vocabulários temáticos	83
Figura 19: Relação total de palavras versus freqüência.....	85

Figura 20: Relação termos versus frequência.....	85
Figura 21: Relação grande área versus termos versus módulo.....	86
Figura 22: Modelo definitivo do Datamart de Vocabulários Temáticos Lattes.....	87

LISTA DE SIGLAS

API	<i>Application Program Interface</i>
C&T	Ciência e Tecnologia
CAPES	Coordenação Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CT-PETRO	Fundo Setorial do Petróleo e Gás Natural
CV	<i>Curriculum Vitae</i>
DMVT	<i>Data Mart</i> de Vocabulários Temáticos
DW	<i>Data Warehouse</i>
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FID	Frequência Inversa de Documentos
FINEP	Financiadora de Estudos e Projetos
FNDCT	Fundo Nacional de Desenvolvimento Científico e Tecnológico
IDF	<i>Inverse Document Frequency</i>
KDD	<i>Knowledge Discovery in Databases</i>
KDT	<i>Knowledge Discovery in Text</i>
MCT	Ministério da Ciência e Tecnologia
MEC	Ministério da Educação e Cultura
OLTP	<i>On-line Transactional Processing</i>
P&D	Pesquisa e Desenvolvimento
PITCE	Política Industrial, Tecnológica e de Comércio Exterior
RI	Recuperação de informação

SAD	Sistemas de Apoio a Decisão
SGVT	Sistema de Geração de Vocabulários Temáticos
SI	Sistemas de Informação
SIG	Sistemas de Informação Gerencial
SPD	Sistema de Processamento de Dados
TFIDF	<i>Term Frequency Inverse Document Frequency</i>
TI	Tecnologia da Informação

GLOSSÁRIO

Ad hoc – frase latina que significa "para este propósito". Geralmente significa uma solução que foi desenvolvida com um propósito específico

Analizador léxico – Sistema capaz de analisar seqüências de caracteres, identificando se partes desta seqüência se encontram dentro de um alfabeto pré-definido

Data mart - Contém dados provenientes do DW, customizados para suportar necessidades analíticas de uma determinada área ou processo de negócio

Data warehouse - Processo de replicação de dados de forma a estruturar e gerar informações, facilitando o processo de análise, consulta e geração de relatórios - armazém de dados

E-mails – Mensagens, normalmente texto, enviado de uma pessoa para outra através do computador

Granularidade – Tamanho mínimo da informação

Hipertexto – Forma de apresentação ou organização de informações escritas, em blocos de texto estão articulados por remissões, de modo que, em lugar de seguir um encadeamento linear e único, o leitor pode formar diversas seqüências associativas, conforme seu interesse.

Homógrafos – duas palavras com significados distintos mas ortografias idênticas

Indexar - Ordenar em forma de índice (palavras, frases, etc.)

Interface – Conjunto de desenhos, gráficos, menus e/ou entradas de dados que possibilitam a interação de um usuário com ferramentas computacionais.

Internet – Rede mundial de computadores de acesso público, cujos principais serviços oferecidos são o correio eletrônico, o *chat* e a Web, a qual é constituída por um conjunto de redes de computadores interconectadas por roteadores que utilizam o protocolo de transmissão TCP/IP. Esta rede foi desenvolvida visando facilitar, num primeiro momento, o livre fluxo de informações entre as universidades. Contudo, com o aumento da variedade de informações disponibilizadas na rede, despertou o interesse da comunidade, inclusive de empresas, visando tanto a troca de informações quanto transações comerciais

Intranet - *internet* particular dentro da organização, a qual evita a entrada de intrusos do mundo exterior. Funciona de forma semelhante à *internet*, e o acesso e a distribuição do conteúdo são feitos internamente

Knowledge Discovery in Databases - representa o processo de transformar dados com baixo nível de conhecimento em alto nível, e ainda enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados

Link – elo de ligação a algum tipo de informação

Microsoft Excel – Programa de editoração de planilha eletrônica

Modelo Espaço Vetorial – Modelo espacial para representação de documentos e seus respectivos termos, comumente utilizado nos sistemas de recuperação de informação; vector space model)

Multimídia – Combinação de diversos formatos de apresentação de informações, como textos, imagens, sons, vídeos, animações, etc., em um único sistema

On-line - termo utilizado para indicar que os dados podem ser acessados direta e imediatamente por um microcomputador ou estação de trabalho

On-line Transactional Processing - Processamento transacional *on-line*

Projeto Jakarta – oferece um conjunto de soluções Java na forma de código aberto fazendo parte da Apache Software Foundation que encoraja um processo de desenvolvimento colaborativo e de desenvolvimento baseado no consenso de uma licença de software aberto

Recuperação de informação (RI) - aplicação computacional de uma tecnologia para aquisição, organização, armazenamento, recuperação e distribuição de informação

Re-parametrização – Processo de gerar novas características como combinação ou transformação de características originais

Site – Conjunto de documentos apresentados ou disponibilizados na Web por um indivíduo, instituição, empresa, etc., o qual pode ser fisicamente acessado por um computador e em endereço específico da rede

Software - sistema operacional cujos recursos permitem que as aplicações sejam processadas com eficiência

Stemming - Técnica de redução de uma palavra para a sua forma raiz.

Stoplist - Lista de stopword

Stopword - Palavras consideradas irrelevantes para propósitos de busca, pois normalmente elas são utilizadas como conectores textuais. Para economizar espaço e tempo, estas palavras são removidas no momento da indexação e então são ignoradas na consulta de busca. Alguns motores de busca permitem incluir uma ou

mais stopword incluindo um sinal de mais antes de cada uma delas

String – uma seqüência de caracteres

Termo - Vocábulo ou locução que denomina um conceito, prévio e rigorosamente definido, peculiar a uma ciência, arte, profissão, ofício

Tesouro - Vocabulário controlado e dinâmico de descritores relacionados de forma semântica e genérica, que cobre de forma extensiva um ramo específico de conhecimento; thesaurus

Vetor de contexto - Representação semântica de palavras e contextos contidos em um texto na forma vetores em um espaço vetorial, onde as dimensões correspondem às palavras; context vector

Web – Recurso ou serviço oferecido na *Internet*, e que consiste num sistema distribuído de acesso a informações, as quais são apresentadas na forma de hipertexto, com elos entre documentos e outros objetos localizados em pontos diversos da Rede

RESUMO

Esta pesquisa propõe um método para a construção de vocabulários controlados temáticos e vetores de contexto, que associam termos a suas frequências, a partir de currículos Lattes, com cálculo de aderência entre vetores. A geração dos vocabulários temáticos é feita por meio da normalização, redução de escopo com *n*-gramas e indexação das palavras-chave dos currículos Lattes. As palavras-chave são associadas às áreas de atuação, formação e produção, considerando a titulação máxima e contabilizando indicadores de frequência e densidade. O cálculo de aderência entre vetores de contexto de currículos e vocabulário utiliza a medida do co-seno. A geração de vocabulários temáticos e o cálculo de aderência fazem parte de um sistema de apoio à decisão de fomento no âmbito dos fundos de apoio à pesquisa. Os vocabulários temáticos geram subsídios para a concepção de editais de chamada de propostas. Os vetores de contexto de pesquisadores e editais podem ser verificados quanto à aderência e, assim, subsidiar o processo de levantamento de possíveis candidatos à avaliador de propostas de financiamento de pesquisa. Os vocabulários temáticos podem subsidiar a construção de tesouros.

Palavras-chaves: Vocabulários controlados; Vetores de contexto; Recuperação de informação.

ABSTRACT

This research proposes a controlled vocabulary construction method and context vectors, which associate terms and its frequencies, from Lattes curriculum vitae (CV), with calculation of adherence among vectors. The generation of controlled vocabulary is made through normalization, scope reduction with n-gram and CV keywords indexation. The keywords are associated to profession, formation and production areas, considering the maximum degree and frequency counting and density indicators. The adherence calculation among CV context vectors uses the cosine measure. The controlled vocabulary generation and the adherence calculation are part of Brazilian sectorial funds fomentation support decision system. The controlled vocabularies generate subsidies for the proposals call proclamations construction. The announcements and CV context vectors can be verified as to its adherence and, to subsidize the rising process of possible candidates to research financing proposals assessor. The controlled vocabularies can subsidize the thesaurus construction.

Keywords: Controlled vocabulary; Context vector; Information retrieval.

CAPÍTULO I – INTRODUÇÃO

1.1 Contextualização

Estudos como de Prusak (1998), Sveiby (1998) e Stewart (1998) destacam a passagem para um período de economia focada nos fatores de produção tradicionais como: trabalho, capital e terra; devido à emergência da sociedade focada no conhecimento e na informação, (DRUCKER, 1995), (NONAKA, 1997) e (DAVENPORT, 1998).

Esta transição tem afetado tanto os países desenvolvidos, quanto os que estão em processo de desenvolvimento. No Brasil, por exemplo, verifica-se o direcionamento de um número cada vez maior de pessoas na busca pela expansão de seus conhecimentos. Rodrigues (1988, p. 29)

considera que a institucionalização da política científica, no Brasil, é expressa pela criação do CNPq (Conselho Nacional de Pesquisa), e da CAPES (Campanha de Aperfeiçoamento de Pessoal de Ensino Superior) 1951. A criação destes dois órgãos foi considerada de fundamental importância para o desenvolvimento da investigação científica e tecnológica no País.

Albuquerque (1982), a fim de destacar a importância da criação do CNPq, observa que somente após a criação do órgão o governo federal passou a apoiar atividades como a pesquisa, permitindo a expansão do potencial científico e

tecnológico do Brasil. Este processo culminou com o aumento da demanda por cursos de pós-graduação, gerando maior número de pesquisadores e conseqüentemente maior demanda por parte do governo tanto em relação a aspectos financeiros quanto à seleção dos projetos encaminhados.

Outro aspecto a ser destacado é o registro pela CAPES entre os anos de 2000 a 2003 de um crescimento de 45,26% no número de mestres e doutores formados neste período, pois nos últimos 10 anos o CNPq aponta um crescimento 244,34%, conforme Figura 1.

	1993	1995	1997	2000	2002
Instituições	99	158	181	224	268
Grupos	4.402	7.271	8.632	11.760	15.158
Pesquisadores (P)	21.541	26.799	34.040	48.781	56.891
Doutores (D)	10.994	14.308	18.724	27.662	34.349
(D)/(P) em %	51	53	55	57	60

Figura 1: Número de instituições, grupos, pesquisadores e doutores.
Fonte: CNPq (2004)

A fim de atender esta demanda foram criados os Fundos de Apoio ao Desenvolvimento Científico e Tecnológico, a partir de 1999, os quais são instrumentos para o financiamento dos projetos de pesquisa, desenvolvimento e inovação. Estes buscam garantir investimentos sólidos e permanentes na pesquisa científica e tecnológica do país.

A criação dos Fundos Setoriais representa o estabelecimento de um padrão diferenciado de financiamento para o setor, sendo um mecanismo inovador de estímulo ao fortalecimento do sistema de C&T nacional (FINEP, 2005). Assim seu objetivo é garantir a estabilidade de recursos para a área e criar um modelo de gestão, com a participação de vários segmentos sociais, além de promover maior

sinergia entre as universidades, centros de pesquisa e o setor produtivo.

Esses recursos, oriundos de contribuições incidentes sobre o faturamento de empresas e/ou sobre o resultado da exploração de recursos naturais pertencentes à União, são alocados no FNDCT. A FINEP, agência responsável pela gestão executiva, atua sob orientação dos Comitês Gestores. Esses Comitês envolvem representantes dos setores produtivo e acadêmico, bem como de diversas instâncias do Governo e definem diretrizes como os planos anuais de investimentos para os Fundos.

O Brasil, atualmente, conta com 16 Fundos Setoriais, aprovados por lei e, juntos no ano de 2004, representaram um acréscimo de R\$ 1,4 bilhões no orçamento da União para C&T - uma ação inovadora e evolutiva da política pública para pesquisa e desenvolvimento (MCT, 2004).

Pode-se destacar o surgimento dos fundos como o processo de privatização de setores da economia, em 1999 com o CT-PETRO e após a publicação dos instrumentos legais regulamentando seu funcionamento. A criação deste serviu como piloto para outros Fundos, e em julho de 2000, o Congresso Nacional, sancionou as Leis que criaram os Fundos Setoriais de: Energia Elétrica, Recursos Hídricos, Transportes, Mineração e Espacial (FINEP, 2005).

Em 2001 foram criados os fundos de: Tecnologia da Informação, Infra-Estrutura e Saúde; assim como também os Fundos de: Agronegócio, Verde-Amarelo, Biotecnologia, Setor Aeronáutico e Telecomunicações. Por fim, em julho de 2004 foram definidas pelo Comitê de Coordenação dos Fundos Setoriais, as Ações Transversais, os quais são programas estratégicos do MCT com ênfase na Política Industrial, Tecnológica e de Comércio Exterior (PITCE) do Governo Federal com utilização de recursos de diversos Fundos Setoriais simultaneamente (MCT, 2004).

Cabe destacar o fato de todos os fundos estarem em operação.

1.2 Problematização

Verifica-se que a introdução dos Fundos Setoriais no sistema de ciência e tecnologia do Brasil implica na injeção de novos recursos nos processos de desenvolvimento científico e tecnológico do País, em especial nas áreas temáticas dos fundos e em intercâmbios entre o setor produtivo, bem como com o sistema científico brasileiro visando à inovação.

Para alcançar tais resultados, a gestão dos fundos deverá ser eficiente e estabelecer prioridades tais como: (a) investimentos, (b) elaboração de planos, (c) programas de desenvolvimento, (d) seleção, (e) avaliação e (f) acompanhamento dos projetos e programas, fornecendo assim subsídios que auxiliem, tanto na identificação de ações, quanto na tomada de decisão.

Neste âmbito a problemática deste estudo está focada em auxiliar os gestores dos Fundos Setoriais no processo de tomada de decisão, a fim de permitir auxiliá-los na distribuição dos investimentos em Ciência, Tecnologia e Inovação.

1.3 Objetivos

1.3.1 Objetivo geral

Construir um mecanismo de geração automática de vocabulários temáticos a partir de repositórios de informação curricular bem como prover mecanismos para o cálculo de aderência utilizando vetores de contexto

1.3.2 Objetivos Específicos

- a. Obter uma base de informação curricular e selecionar a informação relevante para a construção dos vocabulários temáticos;
- b. Projetar um repositório de informações analítica apropriada para extrair a informação relevante para os vocabulários temáticos;
- c. Construção de um sistema de apoio à decisão para a gestão dos editais de projetos.

1.4 Justificativa

A fim de gerar subsídios à tomada de decisão aos gestores dos Fundos Setoriais, utilizar-se-á de informações obtidas a partir da Plataforma Lattes, bem como, da técnica de *data warehouse* (DW) visando à construção de um sistema de apoio à decisão para a gestão dos fundos. O conjunto de ações combina essas técnicas permitindo que a granularidade do DW da Plataforma Lattes seja detalhada ao nível de termos e palavras de consulta dos gestores, através de vocabulários temáticos, bem como possibilita a consulta de currículos que possuam aderência a um edital de um fundo setorial, auxiliando os gestores em suas atividades de programação e avaliação para o fomento.

Este estudo tem a característica de transcender o âmbito dos fundos setoriais, pois o resultado da sistemática desenvolvida, através da geração automática de vocabulários temáticos e da busca textual de currículos com cálculo de aderência, podem ser utilizados como insumos no processo de gestão e análise de projetos.

1.5 Contextualização do trabalho na Engenharia de Produção

Com relação ao problema de pesquisa proposto e ao foco de pesquisa abordado pela Engenharia de Produção, cabe destacar o entendimento comum tanto do *American Institute of Industrial Engineering* quanto da Associação Brasileira de Engenharia de Produção, os quais consideram a necessidade de implementar melhorias aos sistemas produtivos.

Este trabalho traz uma contribuição para os mecanismos de indexação e busca, uma vez que os vocabulários controlados resultantes são elementos importantes em sistemas de indexação e busca, e podem servir de insumo para a construção de outros vocabulários ou tesouros.

Outro ponto a ser destacado refere-se ao fato deste estudo propor um sistema de apoio à decisão, o qual pode auxiliar no processo de tomada de decisão, focando as atividades de programação e avaliação dos Fundos Setoriais.

1.6 Metodologia

A pesquisa realizada é de caráter exploratório, buscando agregar maior conhecimento sobre a linha de pesquisa, desenvolvendo hipóteses a serem testadas e aprofundadas. Segundo Gil (1995, p. 45) a pesquisa exploratória visa “[...] proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explicativo ou a construir hipóteses. [...] tem como objetivo principal o aprimoramento de idéias ou a descoberta de intuições”.

Por desempenhar papel fundamental no desenvolvimento da pesquisa. Lakatos (1990, p. 82) destaca que

[...] o método é o conjunto de atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objeto – conhecimentos válidos e verdadeiros –, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões [...].

Portanto, cabe à metodologia delinear como e onde a pesquisa é realizada, pois nesta etapa são definidos os critérios e os instrumentos utilizados para o desenvolvimento do estudo. Portanto, ao desenvolver uma pesquisa, a metodologia é considerada fator relevante, pois, é por meio dela que o pesquisador decidirá “acerca do alcance de sua investigação, das regras de explicação dos fatos e da validade das generalizações” (GIL, 1995, p.28).

Esta pesquisa, além de possuir caráter exploratório, é de natureza qualitativa, pois se preocupa com a identificação da presença ou da ausência dos

fatores analisados. Com relação aos aspectos qualitativos, Lakatos (1990, p. 103) destaca que “[...] a mudança qualitativa seria a passagem de uma qualidade ou de um estado para outro. O importante é lembrar que a mudança qualitativa não é obra do acaso, pois decorre necessariamente da mudança quantitativa [...]”, por isso, nas pesquisas qualitativas são encontrados fatores quantitativos. Contudo, Fachin (1993) salienta que cabem às pesquisas qualitativas não apenas aspectos mensuráveis, mas principalmente a análise descritiva desses aspectos.

Ainda com relação aos procedimentos técnicos, esta pesquisa enquadra-se no método de estudo de caso, o qual, segundo Gil (1995, p. 78), “[...] é caracterizado pelo estudo profundo e exaustivo de um ou poucos objetos, de maneira que permite o seu amplo e detalhado conhecimento”, proporcionando condições de reunir detalhes e contribuindo para a obtenção de resultados amplos.

Com relação ao método de trabalho para a estruturação de um sistema de apoio a decisão para os Fundos Setoriais foi delineada várias etapas, as quais podem ser verificadas na Figura 2.

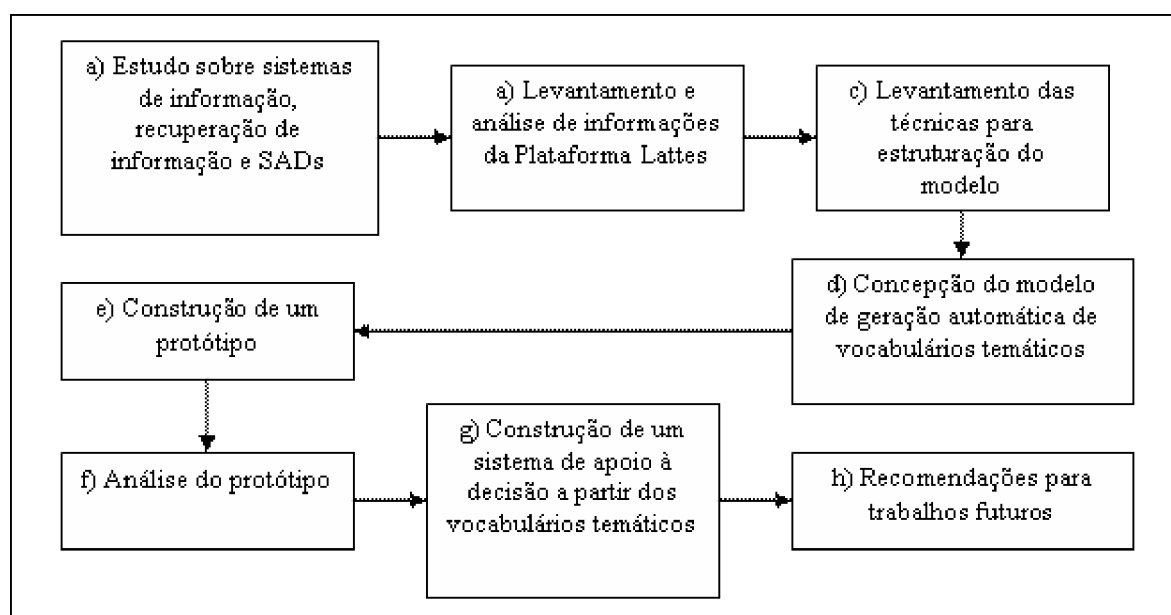


Figura 2: Método para desenvolvimento e implementação do estudo

A fim de promover um melhor entendimento da Figura 2, a seguir é apresentado o detalhamento de cada etapa:

- a. Estudo sobre sistemas de informação, recuperação de informação e SADs – foram identificadas as diversas modalidades de sistemas de informação, bem como delineado os objetivos específicos de cada um dos sistemas, estabelecendo a modalidade mais adequada para estruturação de um modelo de apoio à decisão;
- b. Estudo do modelo de informação da Plataforma Lattes – estudou-se a estrutura de informação da plataforma bem como foram identificadas as unidades de informação do currículo Lattes;
- c. Levantamento das técnicas para estruturação do modelo – em relação à modalidade de sistema de informação a que este trabalho se destina, realizou-se um estudo da técnica de *data warehouse* para criar um modelo de dados dimensional para suportar os vocabulários temáticos e os vetores de contexto a serem indexados em um sistema de busca textual. Estudo dos processos de KDD e KDT para utilização no processo de transformação e carga do modelo dimensional;
- d. Concepção de um modelo de geração automática de vocabulários temáticos – concebendo um modelo de geração automática de vocabulários temáticos através das informações contidas na Plataforma Lattes por meio das técnicas estudadas na etapa “c”;
- e. Construção de um protótipo – a partir da modelagem estruturada na etapa “d” desenvolveu-se o protótipo de parte do sistema;
- f. Análise do protótipo – o protótipo foi analisado, verificando sua

adequação aos requisitos, e a partir disso foram promovidos ajustes, direcionando a construção do sistema;

- g. Levantamento e análise de informações sobre os Fundos Setoriais – foram realizadas pesquisas a fim de explorar e conhecer as características dos fundos, bem como suas metas e dificuldades enfrentadas no processo decisório;
- h. Construção do sistema - o sistema é composto por duas partes principais: (a) vocabulário temático, o qual processa as palavras extraídas dos currículos da plataforma Lattes, gerando vocabulários e vetores de contexto; (b) sistema de busca textual, realiza a indexação dos vetores de contexto gerados na etapa anterior, utilizando um motor de busca capaz de encontrar currículos que possuam aderência com o vetor de contexto de um edital de projeto definido pelo gestor;
- i. Recomendações para trabalhos futuros – são apresentadas recomendações visando comparar os resultados deste trabalho com outras ferramentas de recuperação de informação.

1.7 Estrutura do trabalho

No segundo e terceiro capítulos são definidos os principais conceitos da revisão da literatura, os quais buscam garantir a confiabilidade à análise realizada. Observa-se ainda que no terceiro capítulo é apresentado o modelo em espiral de desenvolvimento de *software*, seguido do sistema de apoio à decisão para a gestão dos fundos setoriais e da construção do SAD, o qual promove a indexação de informações para a busca textual.

No quarto capítulo são apresentados os resultados da pesquisa; e no quinto capítulo seguem algumas conclusões da dissertação, retomando os objetivos específicos da pesquisa e apontando os aspectos relativos ao seu cumprimento bem como algumas recomendações para trabalhos futuros.

CAPÍTULO II – BUSCA E RECUPERAÇÃO DE INFORMAÇÃO

2.1 Considerações iniciais

Este capítulo contempla o referencial teórico considerado imprescindível para o desenvolvimento deste estudo sendo abordado: (a) a extração, recuperação e armazenamento das informações, (b) o modelo de recuperação de informações espaço-vetorial, (c) a indexação e normalização, (d) a indexação de termos, (e) a remoção de palavras irrelevantes (*stopwords*), (e) a normalização morfológica, (f) o cálculo de relevância do termo, (g) a redução de dimensionalidade, (h) a busca e recuperação, e por fim serão abordados (g) os vocabulários.

2.2 Extração, recuperação e armazenamento das informações

Informações sobre os negócios estão em crescente disponibilidade *on-line*, em formato textual, tanto na *Internet*, quanto nas *Intranets*. Assim verifica-se que a falta de informação não se caracteriza mais como um problema, mas sim o emaranhado e a grande quantidade desta. Nesse sentido a falta de ferramentas

capazes de auxiliar na organização e na extração de informações a custo e tempo aceitáveis, são um fator limitante para extração de informações.

Para isso, se faz necessário eliminar a ambigüidade, a qual é uma característica comum na linguagem natural, uma vez que muitas sentenças têm múltiplas interpretações. Como exemplo pode-se destacar a palavra “banco”, a qual pode estar se referindo a uma instituição financeira, a margem de um rio ou ainda a um repositório de dados, sendo que estas diferenças podem ser percebidas em dois momentos: primeiro devido à análise do contexto em que uma palavra aparece e em segundo momento devido ao conhecimento das pessoas que farão uso da informação.

Jackson e Moulinier (2002) destacam a existência de três passos a serem utilizados para auxiliar na extração de informações, uma vez que a maioria dos textos possui um embasamento lingüístico, (a) iniciando pela análise das estruturas gramaticais (sintaxe), (b) passando pela análise do significado (semântica) e (c) terminando no problema do contexto e da linguagem (tratamento pragmático).

Em relação a RI Jackson & Moulinier (2002, p. 26) destacam sua definição “[...] como uma aplicação computacional de uma tecnologia para aquisição, organização, armazenamento, recuperação e distribuição de informação”. A pesquisa associada ao estudo de RI preocupa-se com a melhoria de tecnologia dos mecanismos de busca, incluindo a construção e manutenção de grandes repositórios de informação. Cabe destacar que nos últimos anos têm-se avançado nos conceitos bibliográficos e de busca de textos completos em repositórios de documentos na *Internet*, os quais são associados a: bancos de dados, hipertexto e multimídia.

Na perspectiva do usuário, a RI pode ser redefinida como uma atividade com o propósito de encontrar documentos relevantes, por meio de um mecanismo

de busca, no qual o usuário preencha as informações necessárias e envia uma consulta.

Assim segundo Jackson & Moulinier (2002, p. 28), faz-se necessária a estruturação de um índice, o qual “consiste de uma lista de todas as palavras ocorridas em todos os documentos na coleção”. A Figura 3 exemplifica a estrutura básica de um índice.

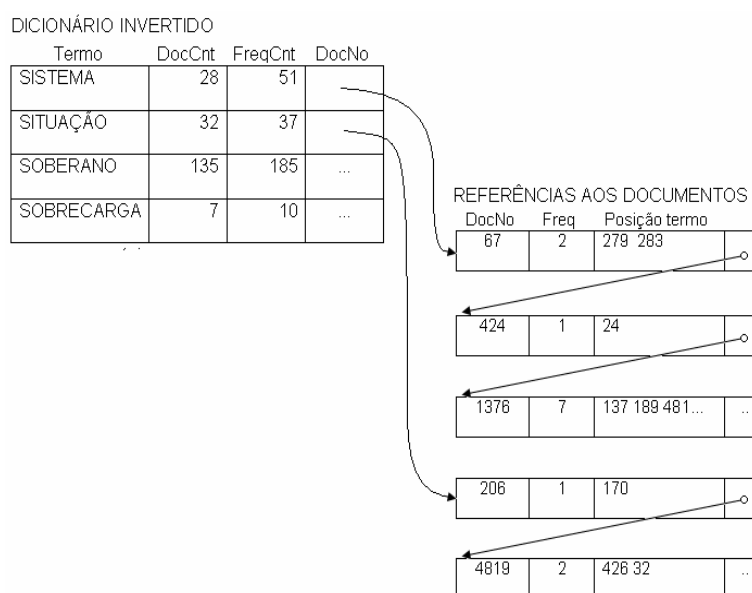


Figura 3: Estrutura básica de um índice de arquivo invertido
Fonte: Jackson & Moulinier (2002, p. 28)

Em relação à Figura 3, verifica-se que para cada termo, armazena-se a seguinte informação: (a) **DocCnt** - contador de documentos, informa em quantos documentos o termo ocorre. Isto permite computar estatísticas úteis, com o propósito de criar uma pontuação de relevância, chamada de FID; (b) **FreqCnt** - contador de frequência total, informa qual é o total de ocorrências de um termo em todos os documentos, esta é uma medida básica de quão comum é o termo; (c) **Freq** - frequência, informa quantas vezes o termo ocorre no documento, este número é um indicador bruto se um documento tem ou não relação com o termo; (d) **Posição termo** - registram as localizações no texto em que as palavras foram encontradas,

estas localizações são registradas por diferentes razões, alguns mecanismos de busca permitem aos usuários realizarem uma consulta informando o número de ocorrências do termo, como exemplo o “Google.com.br”, que usa localizações para gerar partes de texto através destas palavras de contexto, que podem ser bastante efetivos nos resultados dos documentos recuperados.

É necessário destacar que durante o processo de indexação, as palavras devem sofrer algum tipo de tratamento, desse modo as variantes de uma palavra são indexadas apenas uma única vez. Na maioria dos casos, variantes morfológicas das palavras têm interpretações semânticas semelhantes e podem ser consideradas como equivalentes para aplicações de RI.

Segundo Frakes (2003, p. 26) e Kraij (1996, p. 625) uma das formas mais utilizadas para tratar as variantes de uma palavra é a técnica de *stemming*. Esta técnica tenta reduzir uma palavra para sua forma raiz, de modo que termos chaves de uma consulta ou documento sejam representados através da raiz ao invés das palavras originais. Além de diferentes variantes de um termo serem apresentadas apenas sob uma única forma representativa, o *stemming* promove a redução do tamanho de dicionário, ou seja, redução do número de termos distintos necessários para representar um conjunto de documentos.

2.3 Modelo de recuperação de informação espaço-vetorial

A recuperação de informação é um campo de pesquisa cujo desenvolvimento tem sido acelerado nos últimos anos, sendo também criadas

diferentes abordagens. Wives (2002) destaca que estes podem ser: *booleano*, espaço-vetorial, probabilístico, difuso (*fuzzy*), busca direta, aglomerados (*clusters*), lógico e, mais atualmente, contextual ou conceitual. Destes modelos o mais amplamente difundido é o modelo espaço-vetorial.

Neste modelo de recuperação de informação desenvolvido por Salton et al (1975), cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica o grau de importância (denominado peso) deste no documento. Assim, cada documento é representado por um vetor associado que é constituído por pares de elementos na forma {(termo1, peso1), (termo2, peso2), (termo3, peso3)}.

O peso de um termo em um documento pode ser calculado de diversas formas, contudo os métodos de cálculo de peso geralmente se baseiam no número de ocorrências do termo no documento (frequência). Cada elemento do vetor é considerado uma coordenada dimensional. Desta forma, os documentos podem ser colocados em um espaço Euclidiano de “n” dimensões (onde “n” é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso.

Neste espaço Euclidiano, a consulta do usuário também é representada por um vetor. Dessa forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado.

Para isso, se faz uso das medidas de similaridade entre vetores (Vector-Based Matching), as quais podem ser: medida do co-seno, índice de Jaccard, índice de Dice, índice “N”, medidas de sobreposição (*overlap measures*). Dentre estas medidas, uma das mais utilizadas para o processo de comparação vetorial é o co-seno, devido ao seu grau de estabilidade (Egghe, 2002, p. 845). CHIAO (2002, p.4)

realizou um estudo comparativo entre diversos tipos de cálculo, obtendo uma melhor performance através da medida do co-seno, a qual pode ser representada a partir da seguinte fórmula:

$$s(D, Q) = \frac{\sum_k (t_k \times q_k)}{\sqrt{\sum_k (t_k)^2} \times \sqrt{\sum_k (q_k)^2}} \quad (1)$$

O grau de similaridade obtido através da medida do co-seno representa as distâncias entre documentos, ou seja, documentos que possuem os mesmos termos são colocados em uma mesma região do espaço e, em teoria, tratam de um assunto similar (essa característica é que dá o nome de espaço-vetorial ao modelo – Figura 4).

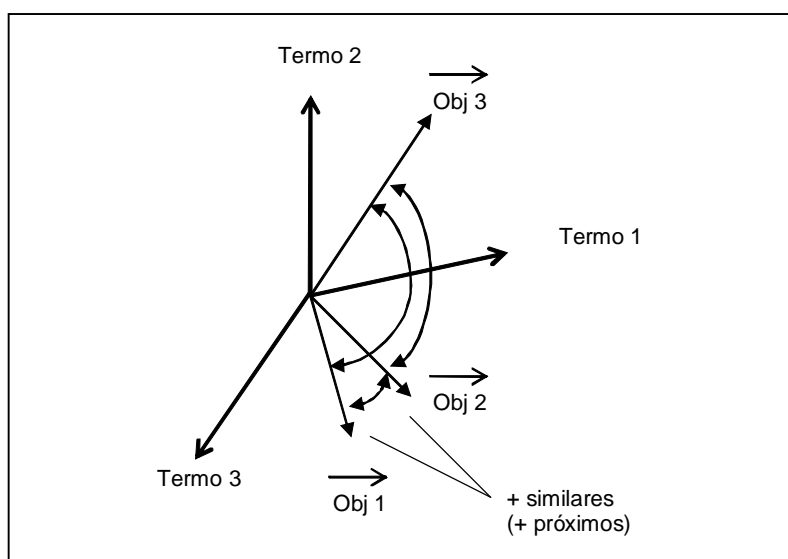


Figura 4: Modelo espaço-vetorial
Fonte Wives (2002)

Em relação ao modelo espaço-vetorial representado na Figura 4, os documentos são representados por vetores. Cada elemento (termo) de um vetor é considerado uma coordenada dimensional. Assim os documentos podem ser colocados no espaço-vetorial de “n” dimensões (onde “n” é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso. A consulta do

usuário também é representada por um vetor. Dessa forma, o vetor de consulta pode ser comparado com os vetores dos documentos e o grau de similaridade com cada um dos vetores pode ser identificado. Os documentos mais próximos à consulta, no espaço-vetorial, ou seja os mais similares, são considerados como relevantes e retornados como resposta ao usuário.

2.4 Indexação e normalização

O processo de indexação automática visa indexar palavras relevantes presentes em uma coleção de documentos e armazená-las em uma estrutura de índice, de forma a facilitar e agilizar a busca de informações. Segundo Wives (2002, p.51) as etapas deste processo podem variar dependendo da situação, contudo as fases normalmente utilizadas são: Identificação de termos, Remoção de *stopwords*, Normalização morfológica (*stemming*), Seleção de termos, Redução de dimensionalidade

Estas técnicas utilizadas para execução destas etapas são descritas na seção identificação de termos.

2.4.1 Identificação de termos

Essa fase refere-se à aplicação de um analisador léxico, de modo a identificar as palavras presentes nos documentos, ignorando os símbolos e caracteres de formatação ou controle de arquivo.

Um dicionário pode ser utilizado para validar as seqüências de caracteres identificadas, de modo a validar sua existência e corrigir possíveis erros ortográficos (Salton, 1983). Um tesaurus ou um dicionário de sinônimos pode auxiliar na normalização do vocabulário, no caso de se utilizar um vocabulário controlado.

Conforme Baeza (1992) e Fox (1992) diversas técnicas adicionais de padronização podem ser aplicadas: a passagem de todos os caracteres para a forma maiúscula (ou minúscula); a substituição de múltiplos espaços e tabulações por um único espaço; a padronização de datas e números; a eliminação de hífen; eliminação de acentuação. É importante ressaltar no caso de uma técnica ser adotada, a mesma também deverá ser utilizada na consulta do usuário.

A padronização gera uma série de vantagens, entretanto a sua utilização pode trazer algumas desvantagens. Um exemplo de desvantagem seria a transformação de caracteres maiúsculos para minúsculos, eliminando a possibilidade de diferenciar substantivos próprios de comuns nas buscas.

2.4.2 Remoção de palavras irrelevantes (stopwords)

Apesar do uso de várias estratégias de peso para modificar a contagem da frequência das palavras e indicar a importância de uma palavra em um documento, o uso de um vocabulário não controlado apresenta diversos problemas.

Muitos estudos têm mostrado que as palavras mais comuns da língua inglesa podem ser consideradas como representação de 50% ou mais de um dado texto qualquer. Normalmente estas palavras não trazem importância relativa ao texto. Tais palavras são os artigos, preposições e outras designações gramaticais de ligação e são designadas como *stopwords*.

Estas palavras trazem dois impactos para sistemas de recuperação de informação. Primeiro devido a sua alta frequência, diminuindo o impacto das frequências das palavras menos comuns. Segundo, elas resultam em processamento de informação com baixo significado quando deixadas no texto. Por estas razões é definida uma *stoplist*, a qual consiste das palavras desconsideradas no processamento. Em geral a *stoplist* contém entre 250 (duzentas e cinquenta) a 300 (trezentas) palavras. Quando um texto é inicialmente processado, cada palavra é comparada com a *stoplist*, e caso seja encontrada, a mesma é ignorada e eliminada da representação do texto.

Em sistemas de recuperação de informação, cada consulta submetida deve ser pré-processada por um sistema de *stoplist* assim como os documentos que foram processados e indexados.

2.4.3 Normalização morfológica

Dependendo do objetivo da indexação, pode ser interessante eliminar as variações morfológicas de uma palavra. As variações morfológicas são eliminadas através da identificação do radical de uma palavra. Para tanto, os prefixos e os sufixos são retirados e os radicais resultantes são adicionados à estrutura de índice. Essa técnica de identificação de radicais é denominada *stemming*, que em inglês significa reduzir uma palavra ao seu radical (ou raiz) (Frakes, 1992; Kraij, 1996).

É possível também identificar os padrões (seqüências de caracteres) que co-ocorrem com frequência nas palavras. Para isso, se deve considerar o texto como uma seqüência de caracteres sem sentido semântico (sem significado) e segmentar essa seqüência em *strings* de tamanho predefinido. Essas *strings* de tamanho fixo são denominadas anagramas (*n*-grams) (Kolwalski, 1997), onde “n” indica o tamanho dos *strings*. Costumam-se trabalhar com bigramas, trigramas e pentagramas. Apesar de ser considerada uma forma de *stemming*, a técnica de anagramas não costuma ser utilizada para fins de recuperação, pois os termos tornam-se incompreensíveis. Outras possíveis aplicações para esta técnica seriam: a criptografia, a detecção de erros ortográficos e a comparação de seqüências de caracteres similares.

2.4.3.1 *Stemming*

Um dos grandes problemas com um vocabulário não controlado surge do fato de uma dada palavra ocorrer em diferentes formas. Por exemplo: “computador”, “computadores”, “computação”, “computacional”; assim como outras palavras que possuem uma mesma forma básica e um conjunto de conceitos correlacionados, este problema implica no fato do usuário ao buscar documentos com a palavra “sistema” deixar de encontrá-los apenas por usar outras formas da mesma palavra.

Uma solução para este problema é a introdução de um algoritmo de *stemming*, o qual retira os prefixos e sufixos da palavra e retorna apenas o seu radical. Caso fosse aplicado o algoritmo ao último exemplo o resultado seria “sistem”. Assim dado documento é representado pela união das várias formas da palavra, resultando numa frequência mais alta e aumentando a significância do termo. Em uma consulta, a utilização do *stemming* assegura a não penalização do usuário devido ao uso de uma palavra específica a qual não ocorra frequentemente.

Vários algoritmos de *stemming* têm sido desenvolvidos com o passar dos anos. Segundo Korfhage (1997, p. 136) os três algoritmos mais conhecidos foram desenvolvidos por Lovins, Porter e Paice. Estes algoritmos utilizam uma abordagem iterativa, ou seja, precisam ser executados várias vezes sobre uma dada palavra, onde em cada execução o algoritmo retira parte do sufixo. Assim uma palavra como “computacionalmente” seria cortada para “computacional” e assim sucessivamente

até chegar a “comput”. A principal distinção entre estes algoritmos se encontra na eficiência de seu código para escolher os sufixos a serem identificados e retirados.

A maioria dos algoritmos de *stemming* utilizados não retira os prefixos das palavras. Korfhage (1997) menciona que isto ocorre porque é difícil identificar qual letra da seqüência é realmente um sufixo ou começo do radical da palavra. Outra preocupação está relacionada com relação ao significado da palavra que normalmente é diferente devido ao uso do prefixo. Isto se aplica ao caso dos prefixos negativos (Ex: **incerto**).

Existem preocupações comuns a qualquer algoritmo de *stemming*. A primeira relacionada à retirada de falsos sufixos de palavras com poucas letras, pode ser resolvida em grande parte pela definição de um tamanho mínimo de palavra a ser processada pelo algoritmo e através de uma pequena lista de exceções. Outro problema em questão está relacionado às palavras cujos radicais são modificados ao sofrerem algum tipo de flexão gramatical, as quais em geral ocorrem como exceções e devem ser tratadas pelo algoritmo.

Finalmente é importante destacar o alto custo de processamento dos algoritmos de *stemming* sobre grandes bases de textos completos. Em contrapartida a representação destes documentos pode chegar a apenas cinco por cento do tamanho original e seu uso mais comum ocorre no pré-processamento das consultas, antes de serem encaminhadas para um sistema de recuperação de informação (Korfhage, 1997).

2.4.3.2 N-gramas

A utilização de n -gramas tem por finalidade reduzir problemas encontrados na comparação de *strings*, tais como masculino/feminino, plural/singular e pequenos erros na ortografia. O n -grama é uma representação vetorial de uma *string* que inclui combinações de n -letras. Segundo Hylton (1996) um vetor n -grama possui um componente para cada combinação possível de n -letras. De um modo geral são utilizados bigramas ou trigramas, ou seja, para compor um vetor n -grama são feitas combinações seqüenciais dois a dois ou três a três, onde “ n ” representa o tamanho da seqüência. Assim, a utilização de “ n ” com grau maior dependerá da aplicação.

Formalmente, um vetor “ \vec{A} ” para uma *string* é definido como:

$$\vec{A} = \{a_{aaa}, a_{aab}, \dots, a_{999}\},$$

onde, “ a_{aaa} ” = número de vezes que “aaa” aparece em “s”.

Considere as seguintes *strings* “Inteligência artificial” e “Inteligência artificial”. Essas *strings* possuem uma pequena diferença resultante de um erro de ortografia. Segundo a definição do vetor n -grama usando “ $n=3$ ” teríamos:

s1 = “Inteligência artificial”

$$\vec{A}_{s1} = \{\text{int,nte,tel,eli,lig,ige,gen,enc,nci,cia,iaa,aar,art,rti,tif,ifi,fic,ici,cia,ial}\}$$

s2 = “Inteligência artificial”

$$\vec{A}_{s_2} = \{\text{int,nte,tel,eli,lig,ige,gen,enc,nca,caa,aar,art,rti,tif,ifi,fi,ici,cia,ial}\}$$

Após a formação dos vetores o algoritmo realiza uma subtração entre

esses vetores utilizando $\|\vec{D}\| = \sqrt{\sum_{i=aaa}^{999} (a_i - b_i)^2}$, onde “ a_i ” representa o “ i th” n -grama de “ s_1 ” e “ b_i ” o “ i th” n -grama de “ s_2 ”.

A comparação entre os trigramas é realizada utilizando um limiar “ T ”, o qual varia linearmente com o número total de trigramas distintos “ n ” nas duas *strings* de entrada. A equação que considera o limiar é dada por:

$$T = 2.486 + 0.025n$$

Para o exemplo, utilizando a equação “ $\|\vec{D}\|$ ”, o resultado seria:

$$\|\vec{D}\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2.$$

Nos dois vetores “ \vec{A}_{s_1} ” e “ \vec{A}_{s_2} ” verifica-se que existem 21 trigramas distintos. Sendo assim,

$$T = 2.486 + 0.025 * 21 = 3.011$$

O limiar permite tolerar erros em *strings* longas, perda de palavras ou erros de ortografia, e possui um bom comportamento em *strings* curtas podendo atuar como um normalizador ideal de palavra quanto ao gênero. Analisando o resultado da subtração e do limiar verifica-se que os dois termos devem ser os mesmos, uma vez que, quanto menor for o valor de “ $\|\vec{D}\|$ ” em relação à “ T ” maior a probabilidade de os termos serem os mesmos.

2.4.4 Cálculo de relevância do termo

Nem todas as palavras presentes num documento possuem a mesma importância. As palavras utilizadas mais freqüentemente (com exceção das *stopwords*) costumam ter um significado mais importante. Palavras constantes em títulos ou em outras estruturas também possuem uma importância maior, pois o autor do documento deve tê-las colocado por considerá-las como sendo muito relevantes e descritivas para fundamentar a sua linha de raciocínio (WIVES, 2002). Assim, os substantivos e complementos também podem ser considerados mais relevantes que os demais termos de uma oração.

Logo, o cálculo de relevância de um termo pode basear-se na freqüência das palavras, na análise estrutural do documento ou na posição sintática de uma palavra. Entretanto, a maioria dos cálculos utiliza a freqüência do termo, ou o número de ocorrências de um termo em um documento.

Além da freqüência do termo é utilizada a freqüência do documento, ou o número de documentos da coleção onde os termos ocorrem. Salton e Buckley (1988) destacam que a medida mais comumente utilizada é a freqüência inversa do documento (IDF), definida como:

$$idf_j = \log\left(\frac{N}{n_j}\right)$$

onde “ N ” é o número de documentos na coleção, e “ n_j ” é o número de documentos em que o termo “ j ” ocorre. O cálculo IDF atribui um baixo peso para os termos com

ocorrência em muitos documentos, por exemplo, classes de palavras como a que contém a preposição “de”, e alto peso para termos que ocorrem apenas em documentos específicos. Uma variação do cálculo IDF adiciona a componente freqüência do termo:

$$tfidf_j = tf_j * \log\left(\frac{N}{n_j}\right)$$

onde “ tf_j ” é a freqüência do termo “ j ” no documento.

Salton e Buckley (1988) testaram várias técnicas para peso dos termos, explorando os efeitos de diferentes combinações da freqüência do termo, IDF, e métodos de normalização. Eles concluíram que não há uma simples abordagem ótima para pesar os termos, mas que o método depende da natureza dos documentos e consultas.

2.4.5 Redução de dimensionalidade

Em modelos estatísticos de representação de documentos ocorre um número elevado de dimensões do vetor de termos. A redução de dimensão visa reduzir o conjunto de termos. Diversas técnicas podem ser usadas para alcançar tal objetivo. No trabalho de Aas e Eikvil (1999), várias delas são citadas. A maioria das técnicas de redução pode ser dividida em duas categorias: seleção de características e re-parametrização.

Seleção de características: visa eliminar termos não informativos dos documentos e construir um conjunto de termos ou radicais que facilite a identificação das classes a qual o documento pertence. Portanto, essa técnica auxilia na

classificação de documentos e reduz a complexidade computacional por meio da redução do número de termos. A idéia principal está baseada no fato de um termo com grande freqüência em diversas categorias ser irrelevante para todas as categorias, e de um termo com rara freqüência, também ser considerado irrelevante.

Um dos métodos utilizados para redução da dimensão baseado na seleção de características é o Corte por Freqüência de Documentos (Document Frequency Thresholding). Neste método é calculada a freqüência de documentos para um determinado termo dentro de um conjunto de documentos sendo removidos os termos com freqüência abaixo de um valor pré-determinado (Yang e Pedersen, 1997). Dessa forma, termos de baixa ocorrência são considerados termos não informativos para a indexação do documento por não possuírem influência no desempenho global.

2.5 Busca e recuperação

A consulta é o modo pelo qual o usuário comunica-se com os sistemas de busca e recuperação de informação. Destaca-se que o usuário especifica sua necessidade de informação, definindo os assuntos dos documentos a serem procurados. Desta forma, ela deve ser especificada corretamente para que os documentos relevantes possam ser recuperados.

O processo de especificação ou formulação de consultas é minucioso e pode ser auxiliado por uma série de ferramentas. Estas são denominadas como ferramentas de auxílio à elaboração de consultas, e guiam o usuário na seleção do vocabulário mais adequado para a sua especificação. As ferramentas de auxílio à

elaboração de consulta mais comuns são os vocabulários temáticos e os tesauros.

2.5.1 Vocabulários

A função dos vocabulários é auxiliar na rápida localização dos termos pesquisados. Para ajudar nesta tarefa existem os sistemas de recuperação de informações, os quais segundo Korfhage (1997) são divididos em quatro etapas:

a) Entradas – documentos (no sentido mais amplo, podendo ser registros de todos os tipos) são obtidas por um centro de informação que implementa o sistema, implicando na existência de critérios e validações na entrada de dados, implicam em um detalhado e apurado conhecimento das necessidades de informações dos usuários a serem atendidos pelo sistema;

b) Indexação – um dos pontos principais de qualquer análise textual, pode ser considerada como uma “representação do documento” ou “vetor de contexto do documento”, onde dois elementos importantes devem ser definidos. Primeiro a descrição física do documento (catalogação descritiva) permitindo que o mesmo seja localizado. E segundo refere-se à escolha dos pontos de acesso (exemplo: autores, títulos), para que a descrição seja passível de ser encontrada;

c) Tradução – pode ou não estar envolvido um vocabulário que consiste em um conjunto de termos, os quais podem ser palavras, frases ou ambos,

d) Saída – a população de usuários a ser servida envia uma série de requisições ao sistema os quais são convenientemente preparados com diversas estratégias de busca, envolvendo também os passos de análise conceitual e tradução da requisição do usuário.

Korfhage (1997) destaca que a análise conceitual procura definir o desejo real do usuário enquanto a tradução trabalha a análise conceitual para uma linguagem que seja compreendida pelo sistema, isto é, uma “representação da requisição” ou “vetor de contexto da requisição” ou ainda “a estratégia de busca” da mesma maneira que um registro do índice é uma representação do documento. A diferença está na representação da requisição, uma vez que pode ocorrer um conjunto de relações lógicas entre os termos indicando como deve ser feita a busca.

Estabelecida à estratégia de busca, a mesma é realizada e as representações de documentos similares a esta estratégia são recuperadas e entregues ao usuário de forma apropriada.

2.5.2 Tesouros

O maior problema a ser considerado é o uso de termos relacionados ou similares. Observa-se que estes não podem ser tratados pelo uso de simples algoritmos para formas variantes de palavras, pois termos similares são normalmente distintos. Por exemplo: uma pessoa pode “enviar” uma carta ou “postar” uma carta, as quais possuem o mesmo sentido. Neste caso os tesouros são utilizados para resolver este tipo de problema. Cabe destacar que, para ser considerado ideal, um tesouro deve conter sinônimos e antônimos para cada palavra, junto com termos longos e curtos, e termos com significados muito próximos. Assim, o tesouro pode ser utilizado durante o processo de armazenamento do documento para controlar o vocabulário, isto é feito substituindo-se cada termo variante com um termo padrão com base no tesouro.

Alternativamente, um tesouro pode ser utilizado durante um processo de consulta, aumentando seu alcance e assegurando a permanência dos documentos relevantes, o que não ocorreria com o uso de um vocabulário restrito.

Uma questão importante a ser analisada é o uso de homógrafos. Neste caso uma das análises: sintática, semântica e pragmática; devem ser aplicadas para determinar as equivalências e diferenças entre tais palavras.

Em sistemas capazes de trabalhar com documentos multimídia ocorre um problema similar com relação aos homônimos, palavras com som semelhante, porém com significados diferentes. As palavras “sessão”, “seção” e “cessão” possuem o mesmo som, contudo “sessão” equivale a uma reunião de pessoas, “seção” é o ato ou efeito de dividir e “cessão” o ato de conceder.

Enquanto o tesouro tradicionalmente foca-se nos termos com significados relacionados, sendo possível desenvolvê-lo baseado na co-ocorrência dos termos, podendo ser muito dependente sobre uma coleção particular de documentos e também baseado nas propriedades semânticas. Entretanto, poderia ser útil saber, por exemplo, se em um conjunto de documentos em particular o termo “recuperação de informação” ocorre em conjunção com o termo “vocabulário controlado”. O conhecimento deste tipo de relação possibilita novos e inesperados caminhos no aumento do alcance de uma busca. Observa-se ainda que, os tesouros baseados em co-ocorrência podem ser gerados automaticamente, com pouca ou nenhuma intervenção humana.

CAPÍTULO III - SISTEMAS DE INFORMAÇÃO

3.1 Considerações iniciais

Este capítulo contempla: a conceituação dos sistemas de informação, bem como a classificação destes sistemas. Em um segundo momento será apresentado as técnicas para a construção de sistemas de apoio à decisão abordando: *data warehouse*, descoberta de Conhecimento em banco de dados e descoberta de conhecimento em Textos, bem como o modelo espiral de desenvolvimento de *software*.

3.2 Conceituação dos sistemas de informação

Os sistemas de informações segundo Laudon & Laudon (1998, p.21), “podem ser tecnicamente definidos como um conjunto de componentes inter-relacionados que coleta (ou recupera), processa, armazena, e distribui informação para a tomada de decisão em uma organização”.

Nas últimas décadas pode-se verificar a evolução pelo qual os sistemas

de informações têm passado. Neste sentido Keen (1996) destaca cronologicamente os seguintes períodos:

- a. 1960 – era do processamento de dados;
- b. 1970 – era dos sistemas de informação;
- c. 1980 – o termo tecnologia da informação passou a ser usado no lugar de informática, processamento de dados e sistema de informação. Isto ocorreu à medida que as telecomunicações passaram a ser o veículo de acesso aos serviços de informática, para os gerenciadores de banco de dados tornaram-se disponíveis nos computadores pessoais e os *softwares* de baixo custo (e sem suporte técnico) invadiram o mercado;
- d. 1990 em diante – era da integração e reestruturação do negócio.

Pode-se destacar que graças aos avanços tecnológicos em *hardware* e *software*, o uso do computador que antes era voltado apenas para agilizar tarefas rotineiras, evolui para uma perspectiva de negócios, intensificando as exigências aos Sistemas de Informação, possibilitando a automação dos processos e do armazenamento das informações, de forma maximizar as atividades das organizações.

Assim Benito (2001, p. 79) ao salientar “a informática: é um campo próspero da ciência, com amplas possibilidades na realização de pesquisas científicas, desenvolvimento de aplicação e outras inúmeras e criativas atividades que poderão ser introduzidas para auxiliar o profissional”, é coerente e oportuna, pois a TI passou a ser percebida como fator de capacitação de novas formas organizacionais, incluindo relacionamentos e processos interorganizacionais.

Assim verificou-se a viabilidade de classificar os sistemas de informações de acordo com sua finalidade, buscando dar suporte às necessidades e metas organizacionais da maneira mais efetiva possível. Na visão de Laudon e Laudon (1998) os sistemas informacionais são compostos de quatro níveis:

- a. níveis operacionais, relacionados com a eficiência de tarefas específicas e com o controle de processos de produção;
- b. nível do conhecimento, relacionado com a criação e com o gerenciamento da informação e os produtos da informação;
- c. nível de controle de gerenciamento, basicamente relacionado com planejamento, controle e monitoramento das atividades operacionais e com o uso dos recursos da organização;
- d. níveis estratégicos, relacionados com o estabelecimento de objetivos organizacionais em longo prazo e a determinação do modo como os recursos e as atividades são controlados.

Com relação aos SI de nível operacional OLTP, Centenaro (2003, p. 20) destaca que “eles representam a automação de funções básicas, transações rotineiras e repetitivas geralmente comuns nos negócios”. Como exemplo, pode-se mencionar: as funções de processamento de folha de pagamento, faturamento, entre outras. Estes SI foram os primeiros sistemas desenvolvidos na grande maioria das organizações, os quais continuam sendo bastante utilizados devido ao seu foco principal estar voltado para os processos. Os principais objetivos destes sistemas são planejar, normalizar, integrar e otimizar os processos.

Os SI de nível de conhecimento, segundo Tait (2000, p. 54) “têm como finalidade apoiar a integração de novos conhecimentos aos negócios e auxiliar a

controlar o fluxo de papéis”.

Os SI de nível gerencial, também denominado SIG, segundo Tait (2000), são responsáveis por transformar dados oriundos de sistemas OLTP e/ou de fontes externas, filtrando e analisando analiticamente os dados, a fim de gerar informações resumidas, obtidas a partir da filtragem e análise de dados altamente detalhados. Os SIGs, direcionados aos gerentes de nível médio (tático), geram e trabalham com um grande número de relatórios e consultas.

Por fim os SI de nível estratégico “são responsáveis por auxiliar os gerentes de alto escalão a planejar suas atividades bem como suportar o planejamento em longo prazo”. (TAIT, 2000, p. 54).

Tait (2000, p. 55) apresenta a classificação dos sistemas em tipos específicos, relacionando-os, tanto aos níveis organizacionais, quanto em relação às informações geradas, conforme Figura 5.

Tipos de sistemas	Nível	Informações Geradas
Suporte a Executivo	Estratégico	Projeções, repostas as perguntas
Apoio a Decisão	Gerencial	Relatórios especiais, análise de decisão, repostas às perguntas.
Informação Gerencial	Gerencial	Sumários e relatórios de exceção
Especialista	Conhecimento	Modelos, gráficos.
Automação de Escritório	Conhecimento	Documentos, e-mail.
Processamento de Transações	Operacional	Relatórios detalhados, listas, sumários.

Figura 5: Inter-relação dos tipos de sistemas com os níveis organizacionais e as informações geradas

Fonte: Adaptado de Tait (2000, p. 55)

Com relação a Figura 5, pode-se destacar que os SAD são considerados como uma evolução dos SIG, sendo os dois voltados ao nível gerencial. Os SAD diferenciam-se dos SIG pela interatividade da manipulação das informações, auxiliando os gerentes em situações não-estruturadas. Outro aspecto em relação a

diferenciação entre os SAD e os SIG refere-se ao foco dos sistemas, conforme demonstra Figura 6.

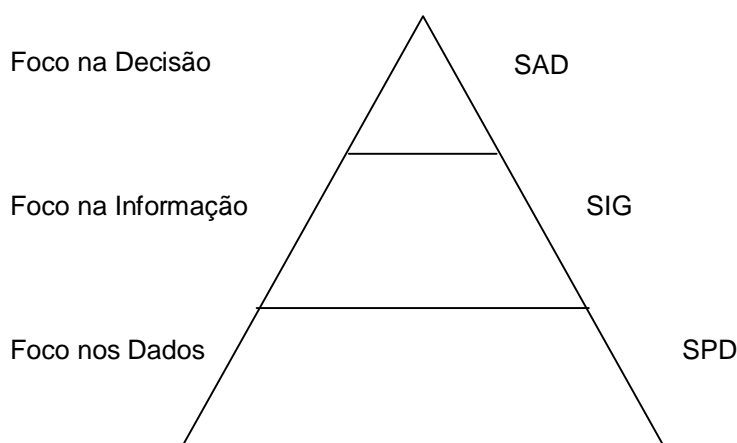


Figura 6: Distinção dos SAD e SIG em relação aos focos
Fonte: GRAHL (1992, p. 9):

Com relação a Figura 6, verifica-se que os SPD são utilizados nos níveis mais baixos da empresa com objetivo de automatizar o trabalho. Como seu foco é nos dados ele possui uma preocupação maior com armazenamento e processamento ao nível operacional. Os SIG encontram-se num nível mais elevado que os SPD, seu foco é nas informações, assim suas atividades estão vinculadas aos sistemas de informações, com ênfase na integração e no planejamento, propiciando consultas e relatórios ao pessoal de nível médio da administração (planejamento tático).

Por fim os SAD são voltados aos níveis organizacionais mais altos. Estes têm foco na decisão, por isso auxiliam os gestores no processo de tomada de decisão, combinando o uso de modelos ou técnicas analíticas com as funções de acesso e recuperação de informações. Estas considerações vão ao encontro das definições de Keen e Morton (1978), que consideram SAD como sistemas computacionais interativos, cujo principal objetivo é dar suporte aos tomadores de

decisão, aumentando a eficácia dos processos decisórios semi-estruturados.

Sprague e Carlson (1982), consideram os SAD como sistemas computacionais interativos, os quais ajudam os responsáveis pela tomada de decisões a utilizarem dados e modelos para resolverem problemas não-estruturados. Mann e Watson (1984), assim como Sprague e Carlson (1982), consideram que os SAD são sistemas interativos que proporcionam ao usuário acesso fácil a modelos decisórios e dados a fim de apoiar as atividades de tomada de decisões semi-estruturadas ou não-estruturadas.

Desta forma, verifica-se que o SAD é a abordagem de sistema de informação que os gestores dos Fundos Setoriais necessitam para ajudá-los no processo decisório.

3.3 Técnicas para a construção de sistemas de apoio à decisão

Por tanto, os SAD, são sistemas de informações ou modelos analíticos projetados para ajudar gerentes e profissionais a tomar decisões mais eficazes. Em geral, os SAD são implementados em computadores pessoais que acessam as bases de dados corporativas. Eles não constituem uma tecnologia específica, mas refletem a ênfase atual à exploração das tecnologias disponíveis para apoiar os executivos, em particular nas análises *ad hoc* e no planejamento.

A palavra “suporte” reforça a importância de que esses sistemas estejam voltados para auxiliar, e não substituir, o julgamento dos executivos. Os SAD mais eficazes associam o acesso a bancos de dados remotos a modelos e ferramentas de

análise. Nos últimos anos as empresas com objetivo de sobreviver, voltam sua atenção a aspectos como eficácia e eficiência, buscando obter vantagem competitiva, neste sentido Benito (2001, p. 92) afirma que

Atualmente a tecnologia de informação é considerada um meio para alcançar vantagem competitiva, o que a torna estratégica na difícil luta pela sobrevivência no mercado. Porém é importante ressaltar que ao levar em consideração este tipo de vantagem, deve-se ter uma atenção especial com relação a segurança destes sistemas que competidores podem tentar copiar, ao elevado custo de manutenção e ao impacto estratégico que, a longo prazo, pode ser incerto, pois na quantidade e velocidade que as informações são geradas hoje, uma informação importante tem um tempo de vida útil cada vez menor.

Neste processo de evolução, foram agregados outros conceitos importantes, “como recurso estratégico, arma estratégica e vantagem competitiva” (TAIT, 2002, p.53), tornando estes sistemas mais do que facilitadores de tarefas rotineiras, mas estratégicos para a sobrevivência e prosperidade das organizações.

Devido a estas características, os SI tornam-se um desafio tanto para os profissionais de TI como para os executivos que irão utilizar-se destas informações. Esse desafio possui alguns princípios fundamentais destacados por Centenaro (2003, p.22), tais como: “*interface* adequada (facilidade de uso e aprendizado), flexibilidade para se adaptar a mudanças que venham a ocorrer em um futuro próximo, e demonstração da informação de forma adequada aos níveis de gerenciamento”.

Além de auxiliar na resolução de problemas de grande complexidade, os quais são mais comuns no cotidiano da alta gerência, os SAD devem possuir uma boa *interface* capaz de transpor as complexidades das consultas que o usuário irá realizar, bem como ter flexibilidade para possíveis modificações futuras (CENTENARO, 2003, p. 22).

Para atender os requisitos citados surgiu o conceito de DW, um sistema paralelo aos sistemas operacionais da empresa, capaz de organizar os dados corporativos subsidiando os gerentes e diretores nas decisões tático-estratégicas. Neste enfoque surge o conceito de DW, o qual visa “atender diferentes necessidades, clientes, estruturas e ritmos que os sistemas operacionais OLTP não atendiam” (KIMBALL, 1998, p. 11).

Segundo Inmon (2002) o DW encontra-se como uma das técnicas mais utilizadas para a construção de SAD. Por meio desta técnica é possível criar sistemas capazes de gerar informações que permitam a análise de fatos para o processo de tomada de decisões. Estas informações são geradas por meio de processamento analítico, as quais permitem a análise de uma grande quantidade de dados em função do tempo, possibilitando através de históricos, a identificação de padrões, tendências e perfis em problemas de grande complexidade. Nesta linha de evolução, tem-se um novo paradigma de SI capaz de gerar informações transversais aos cortes de informações verticais (por setor) obtidas através dos DW.

3.3.1 *Data warehouse* (DW)

Segundo Adelman & Lebaron (1997), o Ambiente de Apoio à Decisão é constituído pelos Sistemas de Apoio à Decisão e Sistemas de Informações Estratégicas. Estes sistemas promovem apoio à decisão, utilizando ferramentas e métodos de análise sobre os dados coletados, são desenvolvidos com intuito de fornecer suporte aos gerentes e pessoas responsáveis pela tomada de decisão.

Permitindo assim uma visualização de vários aspectos do problema a ser analisado.

A multiplicação dos Sistemas de Apoio à Decisão resulta em dificuldades no controle e gerenciamento de dados. Nesse contexto o *Data Warehouse* é uma combinação de conceitos destinados ao suporte da decisão, com o propósito de habilitar o usuário final a tomar decisões melhores e mais rápidas.

Kimball (1998) define DW como o processo de replicação de dados de forma a estruturar e gerar informações, facilitando o processo de análise, consulta e geração de relatórios. Kimball (1998) também o define como recurso de apresentação de consultas de dados empresariais, de um modelo de dados dimensional, tendo como princípio garantir a performance e a legibilidade do modelo por parte dos profissionais de TI, facilitando futuras evoluções do DW.

Na concepção de INMON (2002) o DW é definido como um banco de dados baseado em assuntos, integrado, não-volátil, variável em relação ao tempo e que possui um uso importante no processo de tomada de decisões. O ambiente de *Data Warehouse* passou a funcionar como fonte de dados para os Sistemas de Apoio à Decisão. As preocupações sobre onde buscar as informações e quando carregá-las ficam sobre responsabilidade do DW, permitindo aos Sistemas de Apoio à Decisão se preocuparem apenas com as questões sobre o processamento de consultas.

3.3.2 *Knowledge Discovery in Database (KDD)*

Fayyad (1996, p. 40) define o KDD como o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados, os quais são o conjunto de elementos de um fato (por exemplo, registros de vendas de produtos em um supermercado).

Observa-se o processo não trivial quando alguns tipos de buscas ou inferências estão envolvidas, ou seja, uma computação indireta de quantidades pré-definidas como o computo do valor médio de um conjunto de números. Outro aspecto a ser elucidado refere-se aos padrões válidos, vinculados ao KDD, como sendo expressões em alguma linguagem descrevendo um subconjunto de dados ou um modelo aplicável ao subconjunto.

Goebel & Gruenwald (1999), destacam que o termo KDD é utilizado para representar o processo de transformar dados com baixo nível de conhecimento em alto nível de conhecimento. O processo de KDD pode ser observado a partir da Figura 7.

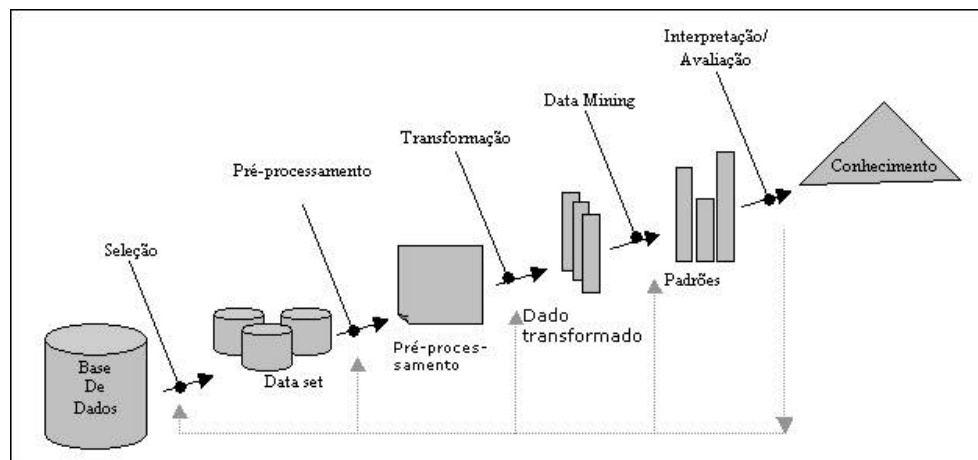


Figura 7: Uma visão dos passos que compõem o processo KDD
Fonte: Fayyad (1996, p. 41)

Cabe destacar que o modelo apresentado na Figura 7 combina métodos e ferramentas das seguintes áreas: aprendizagem de máquina, estatística, banco de dados, sistemas especialistas e visualização de dados (CRATOCHVIL, 1999).

Com relação às técnicas de mineração de dados, estas podem ser aplicadas sobre bancos de dados operacionais, sobre DW ou *Data Marts*, nestes dois últimos casos haverá menos dificuldade para a aplicação de uma técnica de mineração de dados, pois os dados normalmente são preparados antes de serem armazenados no DW ou *data mart* (Dias, 2001). Pode ainda ser aplicada sobre um *data set*, sendo este definido como um “banco de dados”, contendo apenas o conjunto de dados específico para um tipo de investigação a ser realizada.

3.3.3 *Knowledge Discovery in Text* (KDT)

A Descoberta de conhecimento em textos (*Knowledge Discovery in Text* – KDT) pode ser entendida como a aplicação de técnicas de KDD sobre dados extraídos de textos (FELDMAN & HIRSH, 1997). Cabe destacar que o KDT não inclui somente a aplicação das técnicas tradicionais de KDD, mas também qualquer técnica nova ou antiga a ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Deste modo a KDT vem solucionar grande parte dos problemas relacionados à busca, recuperação e análise de informações.

De modo geral o processo de descoberta é composto pelas seguintes etapas: (a) **definição de objetivos**, refere-se à compreensão do domínio, ou seja a identificação de “o quê deve” ou “o quê pode” ser descoberto; (b) **seleção de um subconjunto de dados**, pois nem todas as informações disponíveis são utilizadas, (c) **pré-processamento ou limpeza dos dados**, visa remover ruídos e preparar o dados, (d) **redução ou projeção dos dados**, consiste da escolha de características relevantes para a análise, visto que dependendo do objetivo da análise, determinadas partes podem ser mais importantes do que outras, (e) **escolha da técnica, método ou tarefa de mineração**, como existem vários métodos de descoberta deve-se optar pelo método mais adequado; (f) **Mineração**, consiste da aplicação dos métodos escolhidos; (g) **Interpretação dos resultados**, refere-se ao *feedback* de retro-alimentação do sistema, podendo retornar aos passos anteriores; e (h) **Consolidação do conhecimento descoberto**, vinculado a aplicação prática

do mesmo ou seja relacionado ao uso pelo usuários das informações geradas (JACKSON & MOULINIER, 2002).

Dixon (1997) sugere desenvolver o KDT nos seguintes moldes: (a) **recuperação de informações**, localiza e recupera textos relevantes para “o quê” o usuário necessita descobrir; (b) **extração de Informação**, identifica itens (características, palavras) relevantes nos documentos, (c) **mineração**, aplica técnicas ou métodos de mineração que identifique padrões e relacionamento entre os dados; (d) **interpretação**, interpreta e aplica os padrões e relacionamentos ente os dados identificados.

Assim, verifica-se que a Mineração de Textos ou Descoberta de Conhecimento a partir de Textos refere-se basicamente ao fato de se extrair padrões relevantes e não triviais utilizando-se de bases não estruturadas, sendo em primeira análise vista como uma extensão da Mineração de Dados tradicional. Neste caso a mineração de textos é considerada a versão textual para *data mining*, e consiste de processamento de linguagem natural para extrair conceitos a partir de textos, análise estatística para recuperar padrões entre os conceitos e visualização para permitir análises interativas.

3.4 Modelo Espiral de Desenvolvimento de Software

Na perspectiva de Pressman (1995) o paradigma do modelo espiral é a abordagem mais realística no desenvolvimento de sistemas e de *softwares*, capacitando o desenvolvedor e o cliente a entender e reagir aos riscos em cada

etapa evolutiva. A Figura 8 destaca as quatro principais atividades envolvidas no processo: (a) Planejamento: determinação de objetivos, alternativas e restrições, (b) Análise dos riscos: análise de alternativas e identificação / resolução dos riscos, (c) Engenharia: desenvolvimento do produto na etapa seguinte, (d) Avaliação feita pelo cliente: avaliação dos resultados do desenvolvimento.

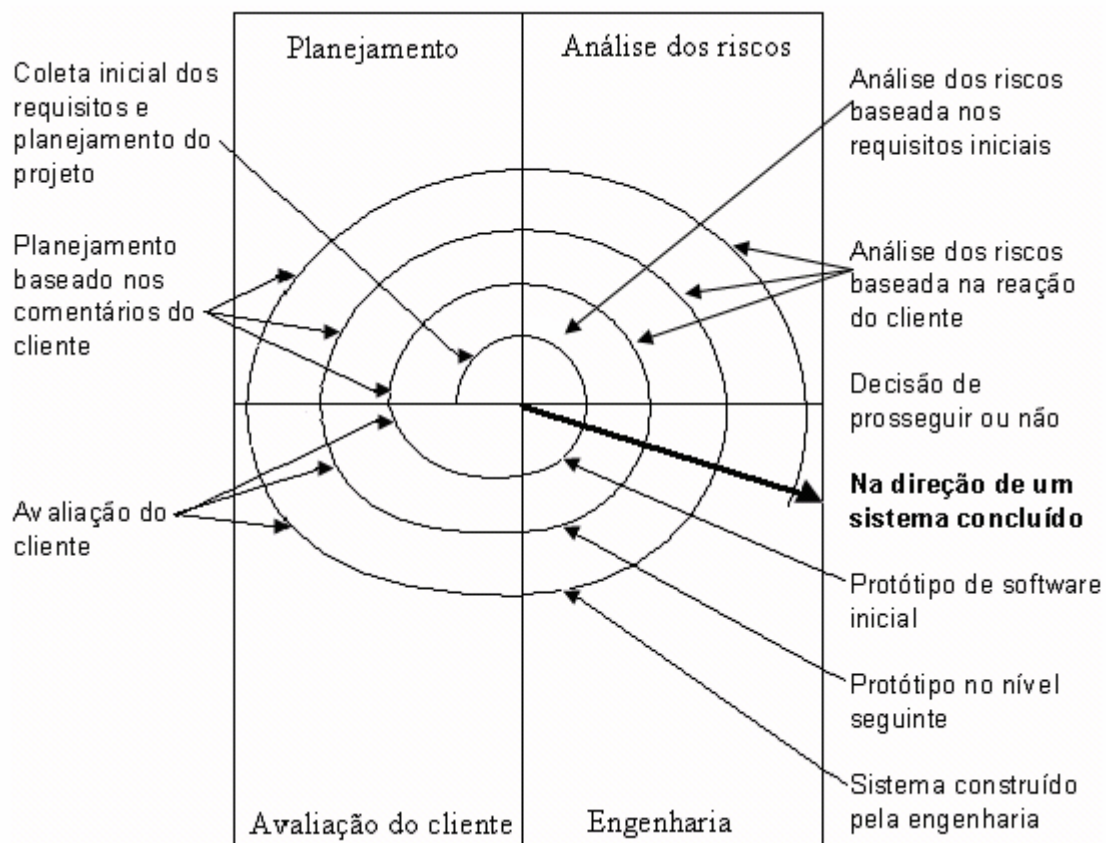


Figura 8: Modelo espiral de desenvolvimento de software
Fonte: Pressman (1995)

O modelo espiral, Figura 8, abrange as melhores características tanto do ciclo de vida clássico quanto da prototipação, acrescentando um novo elemento, a análise de riscos (PRESSMAN APUD BOEHM, 1995). A prototipação é utilizada como um mecanismo de redução de riscos, podendo ser aplicado em qualquer etapa da evolução do produto. Observa-se que no ciclo de vida clássico, são utilizados os

passos sistemáticos, mas incorporando-os a uma estrutura interativa refletindo de modo mais adequado às condições práticas para aplicação. Ao passo que o modelo espiral exige uma consideração direta dos riscos técnicos em todas as etapas do projeto e quando adequadamente aplicado, deve reduzir os riscos antes deles se tornarem um sério problema.

Analisando-se as características do objeto de estudo deste trabalho, como complexidade e requisitos abrangentes, verificou-se a possibilidade da utilização do modelo espiral, sendo esta uma estratégia a ser seguida.

CAPÍTULO IV – GERAÇÃO DE VOCABULÁRIOS A PARTIR DA PLATAFORMA LATTES

4.1 Considerações iniciais

Este capítulo tem por objetivo apresentar o fomento e planejamento em C&T, a Plataforma Lattes, o modelo de sistema de apoio à decisão, os vocabulários temáticos Lattes, a busca textual com o cálculo de aderência, bem como os resultados obtidos com a aplicação do modelo.

4.2 Fomento e planejamento em C&T

Uma das principais questões envolvidas com o Fomento e Planejamento em C&T referem-se à pré-análise da demanda em C&T para os editais nos quais são materializadas as ações de planejamento e fomento do governo.

Para tanto, se deve apresentar o que é um Edital, qual a sua finalidade e quais os benefícios que a busca a partir da descrição do edital trará para o processo de C&T. Cumpre esclarecer que os benefícios consistem em conhecer a oferta de

competência nos campos afetos ao edital. Caso exista muitos CVs, significa que a oferta é farta, assim o decisor pode simular os números da demanda existente em relação à capacidade de atendimento da demanda, caso haja poucos CVs, pode-se prever a necessidade fomentar a formação e a pesquisa nas descritas no edital.

Por tanto o contexto da C&T, fomento pode ser entendido como um mecanismo que visa aumentar a qualificação, o aperfeiçoamento, a especialização e a formação de pesquisadores ligados à ciência e tecnologia, através da concessão de recursos financeiros, materiais ou logísticos. Esse mecanismo é essencial para as atividades de P&D e para a formação de uma ciência básica forte que seja capaz de oferecer condições ao desenvolvimento de projetos, bens e serviços.

4.2.1 Fomento em C&T no Brasil

No Brasil, os investimentos em C&T são quase totalmente financiados pelo Governo Federal, sendo CNPq e CAPES as instituições especializadas que mais investem nessa atividade (MCT, 2004). Ressalta-se também a atuação da FINEP, através do FNDCT, responsável pela infra-estrutura laboratorial de P&D, bem como as agências estaduais, com destaque para a FAPESP, e as empresas estatais.

Entretanto, alguns pontos importantes devem ser considerados, como: (a) a baixa participação da iniciativa privada nos investimentos em C&T e (b) o aumento do número de pesquisadores no país devido ao financiamento das agências de fomento.

Assim uma política de investimentos em desenvolvimento de C&T visando projeções futuras tanto em níveis internos quanto perante a comunidade científica deve estar baseada na definição de áreas prioritárias tidas como ciências básicas, na manutenção e no aumento do vínculo entre as universidades e a iniciativa privada, além do aumento no número de pesquisadores no país.

Krieger e Galembeck (1996) destacam que para o aumento do número de pesquisadores no país são necessárias algumas políticas específicas como: (a) eficiência do sistema educacional, expandindo o ingresso de alunos à universidade; (b) estímulo aos alunos visando a orientação desses para profissões ligadas à C&T; (c) parcerias entre as universidades e a iniciativa privada, objetivando o desenvolvimento de inovações científicas e tecnológicas e promovendo a indústria nacional; (d) recuperação da universidade pública como base da ciência nacional e incentivos às universidades privadas.

A avaliação de C&T é um processo vinculado ao fomento, ou seja, a análise visa produzir subsídios para a tomada de decisão determinando níveis de recursos e quem deve recebê-los. Assim, se faz necessário disponibilizar critérios mais elaborados, levando em consideração a natureza do financiamento.

4.3 Plataforma Lattes

Com a decisão de ampliar o projeto de integração dos sistemas de informações surgiu a Plataforma Lattes, resultado do esforço conjunto de vários órgãos, entre eles MCT, CNPq, FINEP e CAPES/MEC.

Observa-se que a Plataforma Lattes procurou atender a alguns objetivos básicos: (a) implantar alterações e ajustes no sistema de currículos, solicitados pelos consultores *testers*, que entre março e abril de 1999 avaliaram o produto cujo conteúdo ajudaram a definir no ano anterior; (b) ampliar o conjunto de informações e adaptá-las de forma a permitir a adesão da CAPES ao projeto de integração dos sistemas, prevista para o segundo semestre de 1999; (c) estabelecer critérios de avaliação da qualidade das informações coletadas junto à comunidade científica, (c) integrar bases de dados, melhorando o fluxo de informações para pesquisadores, instituições, agências de fomento e órgãos do governo (CNPq, 2004).

Cumprе esclarecer que a Plataforma Lattes é composta de um conjunto de sistemas, os quais promovem suporte à captação e manutenção de dados curriculares dos pesquisadores, dividindo-se em vários sistemas responsáveis, desde o preenchimento dos dados curriculares pelo pesquisador, por meio do Sistema de Currículo Lattes, passando pelos sistemas de recepção dos dados e os sistemas de controle dentro da agência do CNPq. De modo geral, a plataforma fornece subsídios ao incremento e manutenção da base de dados curriculares do CNPq.

4.4 Fundos setoriais

Os Fundos Setoriais segundo o MCT (2004) são os pilares da montagem de um novo padrão de financiamento de Ciência e Tecnologia e de uma nova política de desenvolvimento econômico para o Brasil. Este tem metas vinculadas a diversos pontos de vistas, sendo: (a) gestão de C&T, (b) desenvolvimento econômico nacional, (c) planejamento e coordenação de ações.

É válido destacar que os Fundos não representam apenas uma modalidade diferenciada de financiamento para a pesquisa acadêmica, mas o fato deles terem como foco o desenvolvimento econômico do país e também o tecnológico. O MCT destaca ainda que "os Fundos são vocacionados, atuando em toda a cadeia do conhecimento, desde a ciência básica até as áreas mais diretamente vinculadas a cada setor produtivo, com repercussões sobre o ser humano e sobre o meio ambiente" (FINEP, 2004). Estes visam financiar: encontros e congressos; publicações; auxílios individuais; infra-estrutura de pesquisa; bolsas de formação e de fomento tecnológico; projetos cooperativos entre universidades e empresas; redes cooperativas entre entidades de pesquisas; entre outros; sendo que para obtenção dos financiamentos os solicitantes devem estar vinculados a uma das áreas de pesquisa dos Fundos

Em um primeiro momento os Fundos Setoriais eram financiados por meio de apoio estatal (Orçamento da União); destinavam-se quase exclusivamente às Universidades; e seu controle e acompanhamento feito por agências públicas

criadas segundo a lógica do Estado Desenvolvimentista, sendo a partir dos anos 50 gerido pelo CNPq, CAPES e FINEP.

Assim salienta-se a importância deste estudo uma vez que ele visa gerar subsídios à tomada de decisão aos gestores dos Fundos Setoriais, propondo a integração de modelos de engenharia de software, utilizando as técnicas de *data warehouse*, KDD, KDT, vocabulários controlados e vetores de contexto para a construção de um sistema de apoio a decisão.

4.5 Modelo de sistema de apoio à decisão

Conforme foi abordado no Capítulo 2, os Fundos Setoriais de C&T, são instrumentos para o financiamento dos projetos de pesquisa, desenvolvimento e inovação no País. Este financiamento é feito por meio de recursos próprios, sendo geridos pela FINEP, sob orientação dos Comitês Gestores, os quais definem diretrizes e planos anuais de investimentos para os Fundos.

Através do levantamento teórico das técnicas de DW, RI e vetor de contexto, apoiado por uma metodologia híbrida de desenvolvimento, relacionando o modelo proposto por Presman (1995) às fases iniciais do KDD e do KDT, foi desenvolvido um modelo de apoio à decisão. Este modelo, descrito na Figura 9, visa atender os comitês gestores de C&T, de acordo com os objetivos deste trabalho.

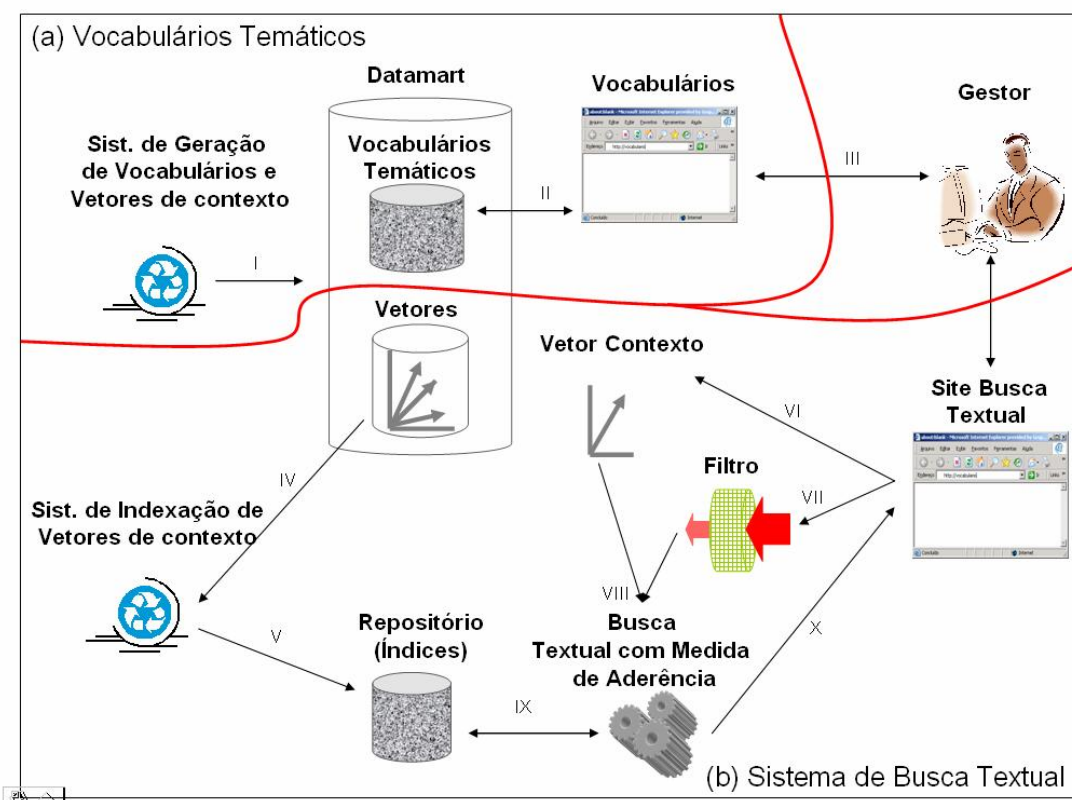


Figura 9: Modelo do sistema de apoio à decisão

Como pode ser observado a partir da Figura 9 o modelo apresenta duas partes principais:

(a) vocabulário temático – composto por um Sistema que processa as palavras extraídas dos currículos da plataforma Lattes, gerando vocabulários e vetores de contexto, armazenando-os em um *data mart* (I), estabelecendo um *site* de consulta de termos por área (II) através do qual o gestor pode gerar os vocabulários (III). Os vetores de contexto gerados nesta etapa são indexados pela segunda parte do modelo

(b) sistema de busca textual – os vetores de contexto armazenados no *data mart* são lidos (IV) por um sistema que os indexa em um repositório de índices (V) para o *site* de busca textual, através deste *site* o gestor defini termos e suas respectivas importâncias (vetor de contexto - VI) bem como alguns filtros para delimitar a consulta (VII), estas duas informações são enviadas ao mecanismo de

busca textual (VIII) o qual efetua a busca no repositório de índices e também um cálculo de aderência comparando cada registro com o vetor de contexto da consulta (IX), o resultado é então retornado ao gestor.

Em uma possível aplicação o gestor pode utilizar o sistema de busca textual para encontrar currículos que sejam aderentes a um vetor de contexto que represente um edital de projeto, este vetor é construído através de uma relação de termos e suas respectivas importâncias. Na *interface* do *site* de busca também são definidos filtros para a busca, de forma a delimitar o escopo. O vocabulário temático Lattes pode ser utilizado pelo gestor, como ferramenta de auxílio na definição dos termos do edital para a busca textual.

4.5.1 Vocabulários temáticos Lattes

O modelo concebido para auxiliar os gestores no processo decisório de C&T foi estruturado considerando as informações e estrutura pré-existentes do DW da Plataforma Lattes, conforme Figura 10.

DW da Plataforma Lattes

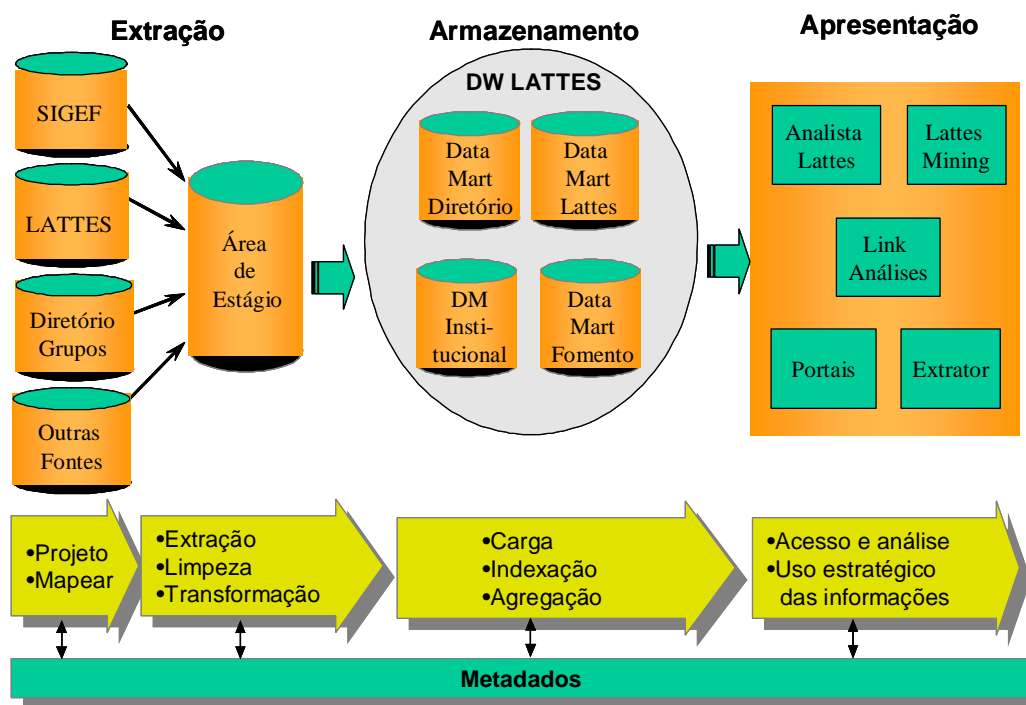


Figura 10: DW da Plataforma Lattes
Fonte: Todesco (2004, p. 7)

Em relação à Figura 10, Todesco et al. (2004, p. 1) destacam que

O *data warehouse* da Plataforma Lattes tem por finalidade estabelecer infra-estrutura de informações e instrumentos para a análise das bases de ciência e tecnologia disponíveis no País, de forma integrada, uniforme e, principalmente, condizente com as demandas dos diferentes atores do sistema nacional de C&T.

Para a implementação do modelo foi feita uma análise e integração ao modelo DW da Plataforma Lattes de mais um elemento denominado *Data Mart* de Vocabulários Temáticos (DMVT), conforme apresenta a Figura 11.

DW da Plataforma Lattes

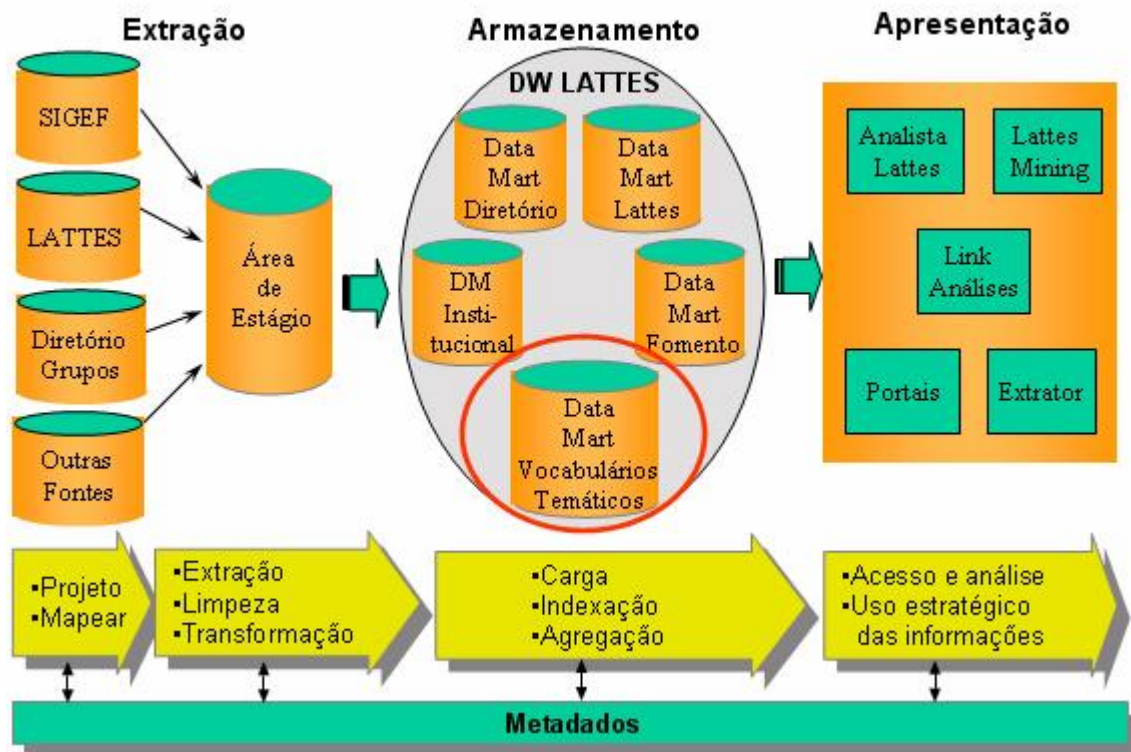


Figura 11: Integração do DMVT ao DW da plataforma Lattes
 Fonte: Adaptado de Todesco (2004, p. 7)

O DMVT, destacado na Figura 11, integra-se aos outros *Data Marts* pré-existentes, criando uma nova visão informacional dos dados curriculares, em um nível de detalhamento maior. A fim de demonstrar esse grau de detalhamento a Figura 12 apresenta a arquitetura do DMVT.

Data Mart Vocabulários Temáticos

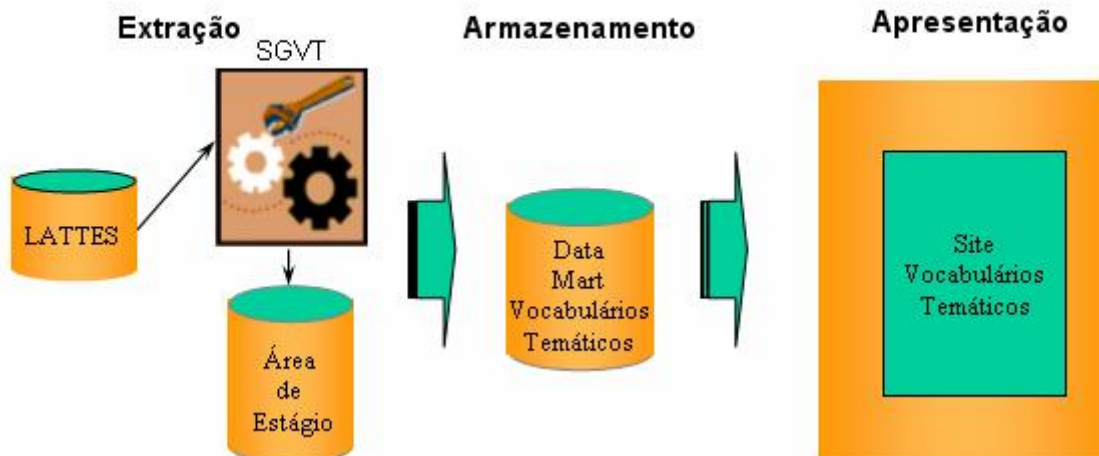


Figura 12: Data mart vocabulários temáticos

Assim, por meio das técnicas discutidas nos Capítulos II e III, foi desenvolvido o SGVT (Sistema de Geração de Vocabulários Temáticos) Figura 13.



Figura 13: Sistema gerador de vocabulários temáticos (SGVT)

O SGVT é encarregado da geração dos vocabulários temáticos, a partir das seguintes tarefas:

- a. Remoção de palavras não significativas através de *stoplists*;
- b. A identificação de termos distintos;
- c. A atribuição das frequências de cada termo, e
- d. Decisão de incluir ou eliminar os termos baseados na frequência.

O sistema SGVT também é responsável por indexar os vetores de termos dos currículos através de uma API chamada Lucene. Esta API é um *software* livre pertencente ao projeto Jakarta, capaz de indexar informações textuais.

Além do SGVT, encarregado pela construção dos vocabulários, foi desenvolvido um *site* para geração dos vocabulários temáticos a partir da definição da grande área, da área, do tipo de informação (atuação, formação e ou produção) vinculada à área e a titulação máxima das pessoas que informaram os termos.

De forma a exemplificar a geração de um determinado vocabulário temático, foi acessada a página dos vocabulários temáticos, procedendo-se com execução dos seguintes passos:

- a. Escolha da grande área: Ciências da Saúde
- b. Escolha da área: Medicina
- c. Seleção da informação vinculada a área: Atuação
- d. Seleção da titulação máxima das pessoas informantes dos termos: Doutores.

A escolha dos filtros utilizados pode ser visualizada a partir da Figura 14.

Ministério da Ciência e Tecnologia Destaque do governo Plataforma Lattes

Página inicial

Vocabulários Temáticos Lattes

1 - Escolher a área do conhecimento para filtrar os currículos:

Grande Área: Ciências da Saúde Instrução

Área: Medicina

2 - Considerar área de:

Atuação Formação Produção (opções)

3 - Filtrar pela titulação máxima:

Doutores Mestres
 Especialistas Graduandos

Gerar Vocabulário

CONHEÇA O CNPq | FOMENTO | LATTES | ATENDIMENTO CNPq | SERVIÇOS CNPq | OPORTUNIDADES MCT | BRASIL.GOV

Figura 14: Site dos vocabulários temáticos - Consulta

Escolhidos os filtros pode ser pressionado o botão “Gerar Vocabulário”, obtendo-se a lista dos termos relacionados (Figura 15).

Ministério da Ciência e Tecnologia Destaque do governo Plataforma Lattes

Página inicial

Vocabulários Temáticos Lattes

Foram encontrados **31104** termos associados à área da **Medicina**
 Mostrando 1 até 10 (página 1 de 3111)

<< anterior || primeira || 1 || última || próxima >>

Termo	Frequência (a)	Áreas nos CVs (b)	Total CVs (c)	Densidade
TRATAMENTO	10628	75	1882	8009.413597
CRIANÇA	9056	74	1290	7832.971965
DIAGNOSTICO	9141	79	1687	7599.239885
EPIDEMIOLOGIA	9336	71	1757	6858.352467
AIDS	7106	70	1283	5825.843386
CARDIOLOGIA	8593	46	880	5276.787836
CIRURGIA	9384	49	1386	5231.214243
CANCER	5258	67	1037	4451.697379
HIV	5024	57	963	3710.792343
MEDICINA	3754	66	914	3266.749987

[Salvar para Excel]

(d) Máxima frequência na área da Medicina: 10628
 (e) Número total de áreas do conhecimento: 99/100,
 onde 99 (e_1) é o número de áreas abrangidas pelos termos selecionados e 100 (e_2) o número total de áreas.
 (f) Número de currículos que contém termos da área: 5026
 (g) Frequência média da área: 45.45
 (h) Densidade: $a * \log_2(f / c) * (b / e_2)$
 * Obs: Clique no título das colunas para classificar o resultado

CONHEÇA O CNPq | FOMENTO | LATTES | ATENDIMENTO CNPq | SERVIÇOS CNPq | OPORTUNIDADES MCT | BRASIL.GOV

Figura 15: Site dos vocabulários temáticos - Resultado

A partir da Figura 15, tem-se um exemplo de um vocabulário gerado a partir da grande área “Ciências da Saúde”, área “Medicina”, considerando apenas a área de atuação, cuja titulação máxima dos informantes do currículo fosse “Doutorado”. Na Figura 15 são vislumbradas várias informações referentes aos termos do vocabulário gerado: o termo, a frequência, as áreas relacionadas, o total de currículos em que o termo foi encontrado e a densidade. Observa-se que a densidade é resultante de um cálculo onde são levadas em consideração as frequências dos termos em relação às outras variáveis conforme a fórmula a seguir:

$$a * \log_{10} (f / c) * (b / e_2)$$

onde

- a. frequência do termo
- b. número total de áreas abrangidas pelo termo
- c. total de currículos em que o termo foi encontrado
- e₂ total de áreas
- f. número total de currículos que possuem termos na área

Parte do cálculo de densidade utiliza a teoria TFIDF (Term Frequency Inverse Document Frequency) de Salton e Buckley (1987):

$$a * \log_{10} (f / c)$$

ao cálculo TFIDF foi adicionado uma componente de peso que leva em consideração a relação entre o total de áreas distintas dos documentos onde o termo foi encontrado e total de áreas distintas existentes:

$$b / e_2$$

através deste componente obtém-se um peso maior para termos que ocorram em currículos com maior abrangência de áreas.

O TFIDF foi escolhido como esquema de cálculo de peso de termos por ser comumente utilizados em sistemas de recuperação de modo a enfatizar as palavras informativas em um texto e reduzir o efeito de “ruído” em outras palavras (YANG e CHUTE, 1994, p. 273).

Em conjunto com as informações mencionadas, a página de resultados contém recursos como: ordenação das informações (ao clicar nos nomes das colunas), *links* para navegação pelas páginas de termos e opção para salvar em uma planilha no formato do *Microsoft Excel*.

Através das informações e dos recursos disponíveis na página de vocabulário gerado pode-se conhecer os termos mais freqüentes: por área, por titulação (doutores, mestres, especialistas e/ou graduandos) e por vínculo dos termos (formação, atuação ou produção) a partir das informações extraídas do currículo Lattes. Estes termos poderão ser utilizados como subsídios para a busca textual com cálculo de aderência curricular.

4.5.2 Busca textual com cálculo de aderência

A maioria dos modelos de recuperação de informação considera a presença de termos em documentos e realizam a busca focando apenas as palavras idênticas. Desta forma, os documentos que possuem as palavras identificadas na consulta são considerados relevantes e os que não apresentam são considerados

irrelevantes por terem uma morfologia diferente. Este tipo de busca é também conhecido como busca binária.

Analisando-se a partir desta lógica, a busca binária é restritiva, pois a linguagem natural possui ambigüidades e incertezas inerentes, causando problemas tanto de sinonímia, onde vários termos podem denotar um mesmo objeto, quanto de polissemia, na qual um termo possui vários significados. Com isso, se um documento de um determinado autor trata do assunto de interesse do usuário, mas um dos dois (autor/usuário) não utiliza os mesmos termos, esse documento não seria encontrado com facilidade.

Existe uma série de ferramentas e técnicas desenvolvidas com o intuito de minimizar os problemas descritos anteriormente. Algumas dessas técnicas sugerem que a compreensão do conteúdo dos documentos e da consulta oferece melhorias significativas ao processo de recuperação, uma vez que o significado dos termos da consulta pode ser identificado.

Com o enfoque destacado anteriormente, a segunda etapa deste estudo de caso aborda o desenvolvimento da busca textual. Este é estruturado contendo as seguintes partes: base indexada de vetores de contexto através da API do Apache Lucene e um motor de busca de vetores de contexto, o qual se baseia no cálculo de similaridade através da medida do co-seno. A indexação dos vetores de contexto foi embutida na aplicação SGVT.

O motor de busca textual e o cálculo de aderência foram implementados em um sistema Web, cuja *interface* pode ser vislumbrada a partir da Figura 16.

Ministério da Ciência e Tecnologia Destques do governo

CNPq PlataformaLattes

Busca textual com cálculo de aderência Página inicial

1 - Escolher a área do conhecimento para filtrar os currículos:

Grande Área: Ciências da Saúde Instrução

Área: Medicina

2 - Considerar área de:

Atuação Formação Producao (opções)

3 - Filtrar pela titulação máxima:

Doutores Mestres

Especialistas Graduandos

4 - Informar os termos da busca e suas respectivas importâncias:

Termo 1	Importancia	Termo 2	Importancia
hiv	Imprescindível	aids	Imprescindível
imunologia	Muito importante		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível
	Imprescindível		Imprescindível

Buscar currículos

CONHEÇA O CNPq | FOMENTO | LATTES | ATENDIMENTO CNPq | SERVIÇOS CNPq | OPORTUNIDADES MCT | BRASIL.GOV

Figura 16: Site de busca textual – Consulta

Através desta *interface* o gestor informa os termos e suas respectivas importâncias. Também são informados alguns tipos de filtros de forma a restringir a busca de currículos, como área de atuação ou titulação máxima.

A partir do momento em que o gestor executa a busca, o sistema mostra uma outra tela, visualizada na Figura 17:

Ministério da Ciência e Tecnologia Destques do governo Plataforma Lattes

Página inicial

Resultados da busca textual com cálculo de aderência

Foram encontrados **1336** ocorrências para a pesquisa realizada.
Mostrando **1** até **11** (página **1** de **122**)

Nome	Titulação Máxima	E-mail	Aderência
Nilo Fernando Rezende Vieira	mestrado	nilo@emescam.br	0.827317732
Daurita Darci de Paiva	doutorado	daurita@uerj.br	0.776094373
Regina Celia de Menezes Succi	doutorado	succi@picture.com.br	0.769056343
Marília de Abreu Silva	mestrado	marilia@predialnet.com.br	0.749836954
Sigrid De Sousa dos Santos	doutorado	ssigrid@hotmail.com	0.738300462
Lauro Ferreira da Silva Pinto Neto	mestrado	laurofp@zaz.com.br	0.738275965
Alex Moura de Barros Nunes	mestrado	alexdebarros@uol.com.br	0.727834359
Marcos Tadeu Nolasco da Silva	doutorado	nolasco@fcm.unicamp.br	0.718787160
Cristina Muccioli	doutorado	cmuccioli@uol.com.br	0.700838853
Jane Margarete Costa	mestrado	carneiom@zipmail.com.br	0.696876500
José Manuel Peixoto Caldas	doutorado		0.695532423

« Anterior 1 2 3 4 5 6 7 8 9 10 11 Próximo »

CONHEÇA O CNPq | FOMENTO | LATTES | ATENDIMENTO CNPq | SERVIÇOS CNPq | OPORTUNIDADES MCT | BRASIL.GOV

Figura 17: Site de busca textual – Resultado

Desta forma, a Figura 17 apresenta o resultado da execução da busca, retornando às pessoas cujos vetores apresentaram aderência ao vetor de contexto da consulta e informando a medida desta aderência através do co-seno, descrito na seção 2.3. As informações podem ser ordenadas através dos títulos das colunas e pode-se acessar as outras páginas de resultado através dos *links* “Anterior”, “Próximo” e dos números entre estes dois *links*.

Em um exemplo de possível utilização o gestor de um fundo setorial poderia definir o vetor da consulta de um edital de projeto e encontrar os currículos das pessoas que possuam uma aderência a este edital. Com estes resultados, o gestor pode preparar uma lista de possíveis candidatos a receber um convite para avaliador de projeto. Contudo é importante destacar o fato desta descrição ser apenas um exemplo de uma possível utilização pelos gestores dos fundos setoriais, sendo que a busca pode ser utilizada por qualquer gestor que trabalhe em atividades de avaliação de editais de projetos.

4.5 Resultados

A base utilizada para geração dos vocabulários Temáticos Lattes, foi extraída da base de currículos Lattes do CNPq do mês de maio de 2003. Nesta base constam 288.017 currículos, com formação, atuação e produção científica associadas a 76 áreas de conhecimento.

Em um primeiro momento foram definidas as partes das quais seriam extraídas as palavras para a geração dos vocabulários: (a) palavras-chave do módulo de formação, (b) atuação e (c) produção científica. A partir da definição foi desenvolvido um modelo de *Data mart* para o vocabulário contemplando as definições iniciais, o qual pode ser verificado a partir da Figura 18.

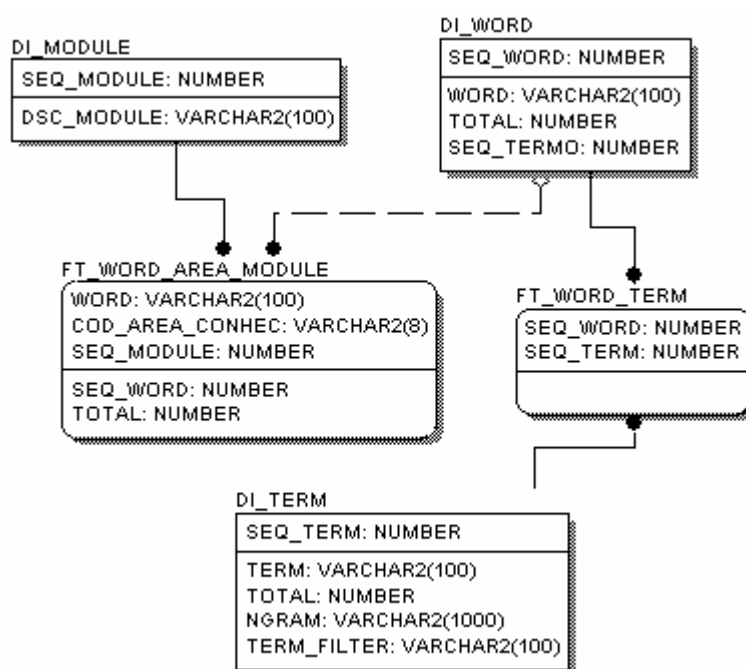


Figura 18: Modelo preliminar da base dos vocabulários temáticos

Com base no modelo do *Data Mart* do vocabulário, foi criado um protótipo para a extração das palavras e criação dos termos. O termo representa a palavra escolhida para indexar e representar outras palavras derivadas, baseado na frequência máxima. Por exemplo, o termo “sistema” irá representar as palavras “sistema” e “sistemas”, ou até mesmo a palavra “sistema”, a qual foi erroneamente informada em um determinado currículo. Para evitar este tipo de problema e diminuir o escopo total de palavras utilizou-se a técnica de remoção de *stopwords* e a técnica de comparação de palavras através de *n*-gramas. Observa-se que não foi necessária a utilização de técnicas de padronização, pois a base do Lattes possui as palavras-chave padronizadas. Devido à complexidade e alto custo de processamento, o protótipo precisou de aproximadamente 11 dias para processar a base de dados extraída do CNPq. A partir da carga do *Data Mart*, foram efetuados alguns cálculos, obtendo-se os seguintes resultados:

- Total de palavras-chave processadas: 13.465.676
- Total de palavras distintas: 1.070.298
- Total de palavras distintas relacionadas a uma determinada área e encontrada em um determinado módulo do currículo: 2.130.210
- Total de termos gerados pela comparação por *n*-grama, havendo uma redução de dimensionalidade da ordem de 25%: 793.261

Ainda em relação ao resultado do processo, podem-se obter outras sumarizações, como a descrita na Figura 19, a qual demonstra a ocorrência de palavras com frequência maior a 15 representando 83,89% de todas as palavras processadas.

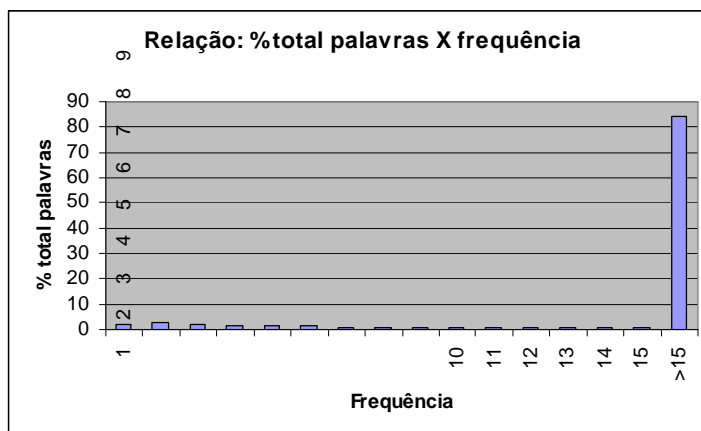


Figura 19: Relação total de palavras versus freqüência

Outra informação importante foi obtida através da Figura 20, a qual demonstra a existência de uma grande ocorrência de termos distintos com freqüência menor a 16, e que os termos com freqüência maior a 15, perfazem um total de 87.868 termos no universo dos 793.261 termos distintos extraídos.

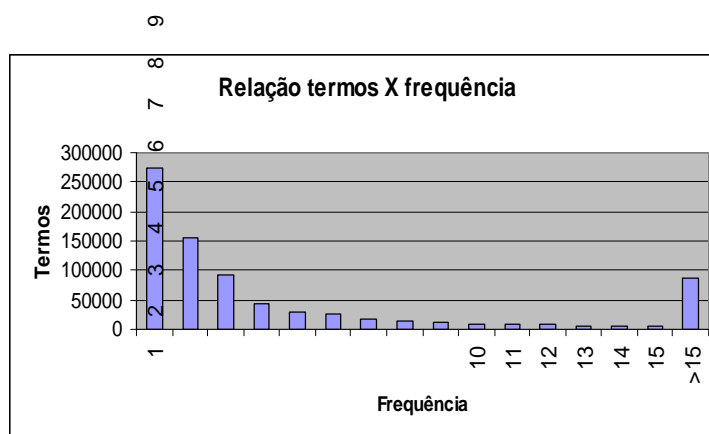


Figura 20: Relação termos versus freqüência

Assim, pode-se concluir que os termos com freqüência maior do que 15, são considerados como mais relevantes para a geração do vocabulário, pois representam 83,89% de todas as palavras da base. Outro ponto importante a ser destacado é que o escopo de termos diminui para 87.868 termos, levando-se em consideração freqüências maiores do que 15, com uma redução de 88,92% do total

de termos gerados. Contudo, devido ao fato da base de currículos ser considerada dinâmica, o escopo pode modificar-se, e um termo tido como irrelevante, devido a sua baixa frequência, com o passar do tempo pode-se tornar relevante. Esta informação foi levada em consideração no desenvolvimento do sistema através da estruturação do SGVT de forma a permitir o reprocessamento das palavras associadas com currículos que foram atualizados ou inseridos na Plataforma Lattes.

Por meio da Figura 21, pode-se chegar a outra conclusão importante em relação aos módulos do currículo, de onde são extraídas as palavras. As palavras-chave de produção representam mais de 98% dos termos gerados, ou seja, a extração a partir do módulo de produção é suficiente para geração do vocabulário.

NOME DA GRANDE ÁREA	TOTAL DE TERMOS	% DO TOTAL DE TERMOS PRODUÇÃO CIENTÍFICA
Ciências Agrárias	34.977	0,991594476
Ciências Biológicas	41.045	0,990717505
Ciências Exatas e da Terra	42.889	0,989717643
Ciências Humanas	31.055	0,989019482
Ciências Sociais Aplicadas	23.420	0,985140905
Ciências da Saúde	35.647	0,991920779
Engenharias	33.770	0,983239562
Linguística, Letras e Artes	12.574	0,986957213

Figura 21: Relação grande área versus termos versus módulo

A partir das considerações da primeira etapa e de uma re-avaliação na estrutura de armazenamento e recuperação de informações dos vocabulários temáticos, foi constatado que a granularidade ao nível de termos não era adequada, e que a relação entre termos e áreas do conhecimento, sem levar em conta a pessoa informante não faria sentido. Então o modelo do *Data Mart* do Vocabulário foi modificado e a granularidade passou a ser os termos por pessoa e não os termos por área (Figura 22)

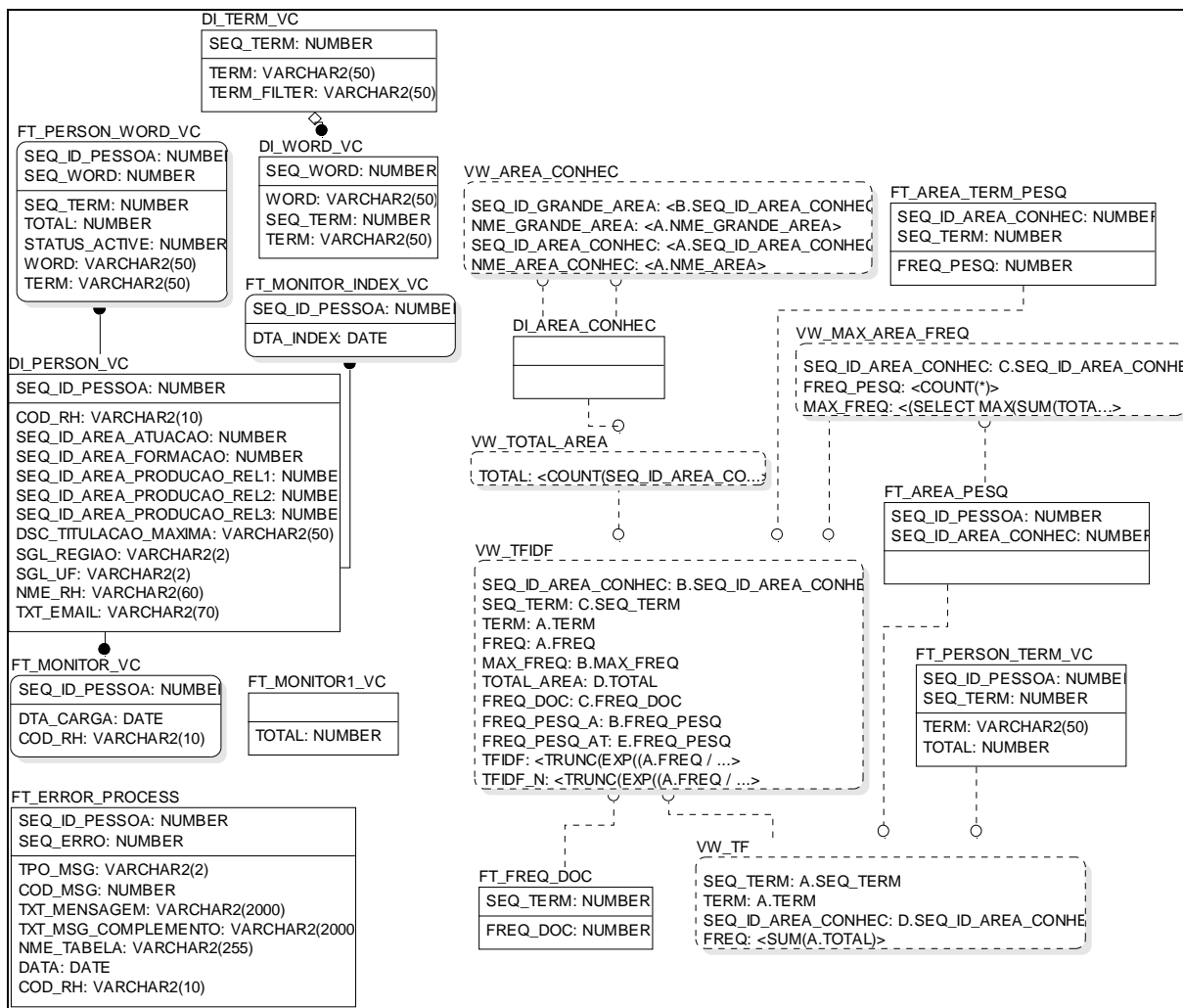


Figura 22: Modelo definitivo do Datamart de Vocabulários Temáticos Lattes

Além da alteração do modelo, na construção do sistema, foi utilizada a técnica de n -gramas para comparação das palavras, de forma a obter uma redução do escopo de termos gerados.

CAPÍTULO V – CONCLUSÕES E RECOMENDAÇÕES

5.1 Conclusões

Com o aumento no número de documentos nas organizações em diversos formatos e conteúdos, tais como memorandos, *e-mails*, relatórios técnicos, artigos, páginas *Web*, entre outros, tornam-se necessárias abordagens que vão além de consultas baseadas em indexação completa ou utilizando campos em bases estruturadas e/ou dimensionais. De um modo geral, um texto tem que ser analisado para que se possa identificar os possíveis termos relevantes para uma consulta bem como seus sinônimos e termos relacionados.

Para a identificação desses possíveis termos relevantes, utiliza-se a extração de características possibilitando que os documentos sejam transformados em vetores de contexto. Em uma visão de C&T essa abordagem pode ser estendida para representar qualquer unidade de análise (exemplo: pesquisadores, grupos e projetos de pesquisa, entre outros). Esses vetores permitem gerar um perfil do documento que facilita tanto a indexação quanto a recuperação de informações relevantes.

Assim considera-se que o objetivo geral deste estudo foi atendido, uma vez que foi desenvolvido o mecanismo de geração automática de vocabulários a partir das informações da Plataforma de currículos Lattes, bem como o sistema de

busca textual com cálculo de aderência curricular. Como resultado do estudo de caso, obteve-se um sistema, o qual dispõe de uma *interface* para o acesso tanto do vocabulário, quanto da busca textual, auxiliando os gestores em suas atividades de programação e avaliação.

Cabe destacar também que os objetivos específicos delineados neste estudo também foram atendidos, por meio da:

- a. Estruturação e integração do *data mart* DMVT para suportar os vocabulários temáticos e os vetores de contexto gerados a partir das palavras-chaves do currículo Lattes;
- b. Implementação do SGVT que integra algoritmos de indexação e comparação de termos, com base na frequência de ocorrência dos mesmos;
- c. Construção de uma base de vetores de contexto curriculares indexadas através do *software* livre Apache Lucene;
- d. Implementação da aplicação de geração dos vocabulários temáticos Lattes a partir da seleção dos filtros: área de atuação, formação e/ou produção bem como da seleção da formação máxima dos informantes curriculares;
- e. Implementação das funcionalidades que permitem aos gestores dos fundos escolherem termos apropriados para buscas textuais de currículos, resultando uma lista de currículos e seu respectivo peso em relação à aderência à consulta, através do cálculo do co-seno.

O mecanismo apresentado nesta dissertação gera vocabulários temáticos que representam o uso de palavras-chave por área do conhecimento com uma

tempestividade que não se obtém pelos modos usuais de produção de vocabulários temáticos, pois a construção parte dos currículos de pesquisadores em seu estado atual. Outro fator importante a ser mencionado deve-se ao fato do mesmo mecanismo captar a prática de uso de palavras-chave em uma área de conhecimento em dado momento. A repetição desta captura permite observar a evolução do vocabulário da área.

5.2 Recomendações

Considerando o trabalho de pesquisa e desenvolvimento realizado, pode-se recomendar o estudo e implementação de outras medidas de aderência como:

- Viabilizar a busca de áreas mais aderentes a um vetor de consulta, possibilitando a visualização destas áreas segundo o grau de aderência;
- Ranking de aderência dos vetores de currículos segundo o vetor de contexto de uma determinada área;
- Fusão do vetor de uma proposta de projeto e seus participantes em um único vetor de forma a possibilitar o cálculo de sua aderência com um determinado edital de projeto.

Outro estudo que merece destaque é a construção de tesouros a partir dos vocabulários temáticos lattes, visto que o custo deste tipo de estrutura normalmente é elevado, quando dependentes exclusivamente do conhecimento de especialistas.

Conforme Capítulo IV verificou-se que através da comparação de termos por *n*-grama houve uma redução de dimensionalidade em torno 15%. Assim, outro possível estudo seria a comparação em termos de performance e redução de dimensionalidade entre a técnica de *n*-grama utilizada na construção dos vocabulários temáticos e a técnica de *stemming*.

REFERÊNCIAS BIBLIOGRÁFICAS

____CNPq – Conselho Nacional de Pesquisa e Desenvolvimento. Disponível em: <<http://www.cnpq.gov.br/>>. Acesso em: 12 Nov. 2004.

____FINEP - Financiadora de Estudos e Projetos. Disponível em: <<http://www.finep.gov.br/>>. Acesso em: 11 Nov. 2004.

____MCT – Ministério da Ciência e Tecnologia. Disponível em: <http://www.mct.gov.br/fontes/artigo.htm>. Acesso em: 11 Nov. 2004.

AAS, K. EIKVIL, L.. **Text categorisation**: A survey. Technical report, Norwegian Computing Center, June 1999.

ADELMAN, S.; LEBARON, M. **Meta Data Standards**. Review Magazine. Dezembro, 1997.

ALBUQUERQUE, Manoel Mauricio de; FENAME. **Atlas histórico escolar**. 8a ed. rev. e atualizada. Rio de Janeiro: FENAME, 1982.

BAEZA-Yates, Ricardo A. **String Searching Algorithms**. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. Information Retrieval: Data Structures & Algorithms. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.219-240.

BENITO, Gladys A. V. **Concepção de um sistema de informação de apoio a supervisão da assistência em enfermagem hospitalar: uma abordagem da ergonomia cognitiva**. Tese – PPGE/UFSC. Florianópolis, 2001. Disponível em: <<http://teses.eps.ufsc.br/Resumo.asp?2955>>. Acesso em: 27/08/2004.

CENTENARO, Antonio César. **Desenvolvimento e implantação de um Data Warehouse corporativo com Data Marts distribuídos em uma cooperativa agroindustrial**. Dissertação – PPGE/UFSC. Florianópolis, 2003.

CHIAO, Yun-Chuang; ZWEIGENBAUM, Pierre. **Looking for candidate translational equivalents in specialized, comparable corpora**. 19th International Conference on Computational Linguistics, Taipei, Taiwan, 2002.

CRATOCHVIL, A. **Data mining techniques in supporting decision making**. Master thesis, Universiteit Leiden, 1999.

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

DIAS, Maria Madalena; PACHECO, Roberto C. S.; **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados**. Universidade Federal de Santa Catarina. Florianópolis, 2001. Tese – PPGE. Disponível em: <<http://teses.eps.ufsc.br/Resumo.asp?1562>>. Acesso em: 14/06/2004.

DIXON, Mark. **An Overview of Document Mining Technology Computer Based Learning Unit**, University of Leeds. October 4, 1997. Disponível em <http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dm.html>

DRUCKER, Peter Ferdinand. **Administrando para o futuro**: os anos 90 e a virada do século. Tradução Nivaldo Montigelli Jr. 5. ed. São Paulo: Pioneira, 1995.

EGGHE; L; MICHEL, C. **Strong similarity measures for ordered sets of documents in information retrieval**. Information Processing & Management, Volume 38, Issue 6, Pages 823-848, November 2002.

FACHIN, Odília. **Fundamentos de metodologia**. São Paulo: Atlas, 1993.

FAYYAD, Usama. **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. November 1996/Vol. 39, No. 11 COMMUNICATIONS OF THE ACM. http://www.ischool.utexas.edu/~i385q-dt/readings/Fayyad_Piatetsky-1996-TheKDD.pdf

FELDMAN R., and HIRSH H. **Mining associations in text in the presence of background knowledge**. In Proceedings of the 2 International Conference on Knowledge Discovery from Databases. 1997.

FOX, Christopher. **Lexical analysis and stoplists**. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. Information Retrieval: Data Structures & Algorithms. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992.

FRAKES, William B. **Introduction to information storage and retrieval systems**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992.

FRAKES, William B.; FOX, Christopher J. **Strength and similarity of affix removal stemming algorithms**. ACM SIGIR Forum, Volume 37 Issue 1, April 2003.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 1995.

GOEBEL, Michael; GRUENWALD, Le. **A survey of data mining and knowledge discovery software tools**. SIGKDD Explorations, ACM. v.1, n.1, p.20-33. 1999. Disponível em: <<http://www.acm.org/sigkdd/explorations>>. Acesso em: 18 Set. 2004.

GRAHL, Everaldo Artur. **Treinamento em sistemas de apoio à decisão baseado na simulação empresarial**. Dissertação, EPS – UFSC. 1992.

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

JACKSON, Peter; MOULINIER Isabelle. **Natural language processing for online applications – text retrieval, extraction and categorization**. Philadelphia, PA, USA: John Benjamins Publishing Company, 2002. Disponível em: <<http://site.ebrary.com/lib/buufsc/Doc?id=10022351>>. Acesso em: 05/12/2004 às 9:45.

KEEN, P. G. W.; MORTON, M. S. Scott. **Decision support systems: an organizational perspective**. Reading Mass: Addison-Wesley, 1978.

KEEN, Peter G.W. **Guia gerencial para a tecnologia da informação**. Editora Campus, Rio de Janeiro – 1996.

KIMBALL, Ralph; REEVES, Laura; ROSS, Margy; THORNTHWAITE, Warren. **The Data Warehouse: lifecycle toolkit**. Wiley Computer Publishing, New York, EUA, 1998.

KOLWALSKI, Gerald. **Information retrieval and storage: Theory and Implementarion**. Boston: Kluwer Academic Publishers, 1997.

KORFHAGE, Robert R. **Information retrieval and storage**. New York: John Wiley & Sons, ISBN 0-471-14338-3, 1997.

KRAIJ, Wessel; POHLMANN, Renée. **Viewing stemming as recall enhancement**. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, August 1996.

KRIEGER, E.; GALEMBECK, F., **A capacitação brasileira para a pesquisa**. In Schwartzman, S.; Bertero, C. O.; Krieger, E. M. et. al. (eds). Ciência e tecnologia no Brasil (vol. 3): A capacitação brasileira para a pesquisa científica e tecnológica. Rio de Janeiro: Editora da Fundação Getúlio Vargas, 1-18, 1996.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 2. ed. São Paulo: Atlas, 1990.

LAUDON, K. C.; LAUDON, J. P. **Sistemas de informação**. Rio de Janeiro: Santuário, 1998.

MANN, R.; WATSON, H. **A contingency model for user involvement in DSS development**. MIS Quarterly, Vol. 8, n.1, p. 27-38, 1984.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação**. Rio de Janeiro: Campus, 1997.

PRESSMAN, Roger S. **Engenharia de software**. São Paulo: Makron Books, 1995.

PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. Rio de Janeiro: Campus, 1998.

RODRIGUES, Alexandre Ferreira. GOELDI, Emilio; **CNPq**. Brasília: Ed. Universidade de Brasília: CNPq, 1982.

SALTON, G. & Buckley, C. **Term-weighting approaches in automatic text retrieval**. Information Processing and Management, 24(5), pp. 513–523, 1988.

SALTON, Gerard; BUCKLEY, Chris. **Term weighting approaches in automatic text retrieval**. Information Processing and Management, Vol. 24, No. 5, pages 513-523. USA, NY: Cornell University, 1987.

SALTON, Gerard; MACGILL, Michael J. **Introduction to Modern Information Retrieval**. New York: McGRAW-Hill, 1983. 448p.

SALTON, Gerard; WOND, A; YANG, C.S. **A vector space model for automatic indexing**. Communications of the ACM, Volume 18, Issue 11: November, 1975.

SPRAGUE, Ralph H.; CARLSON, Eric. **Building effective decision support systems**. Englewood Cliffs: Prentice-Hall, 1982.

STEWART, A. Thomas. **Capital intelectual: a nova vantagem competitiva das empresas**. 4a ed. Rio de Janeiro: Campus, 1998.

SVEIBY, Karl Erik. **A nova riqueza das organizações**. Rio de Janeiro: Campus, 1998.

TAIT, Tânia Fátima Calvi. **Um modelo de arquitetura de sistemas de informação para o setor público: estudo em empresas estatais prestadoras de serviços de informática**. Tese – PPGEP/UFSC. Florianópolis, 2000. Disponível em: <<http://teses.eps.ufsc.br/Resumo.asp?1152>>. Acesso em: 12/11/2004.

TODESCO, José Leomar, et al. **Arquitetura de Data Warehouse da Plataforma Lattes**. Grupo Stela, Universidade Federal de Santa Catarina. Conegov - Conferência Sul-Americana em Ciência e Tecnologia Aplicada ao Governo. Florianópolis, 2004.

WIVES, Leandro Krug. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Exame de Qualificação, Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-graduação em Computação. 2002.

YANG, Y.; PEDERSEN, J. **A Comparative Study on Feature Selection in Text Categorization**. In International Conference on Machine Learning, 1997.

YANG, Yiming; CHUTE, Christopher. **An example-based mapping method text categorization and retrieval**. ACM Transactions on Information Systems, Vol. 12, No 3, July 1994.