EVELINE CRUZ HORA GOMES FERREIRA

***Geração Automática de Metadados:
uma Contribuição para a Web Semântica***

Tese apresentada à Escola Politécnica
da Universidade de São Paulo para
obtenção do Título de Doutor em
Engenharia.

São Paulo
2006

EVELINE CRUZ HORA GOMES FERREIRA

***Geração Automática de Metadados:
uma Contribuição para a Web Semântica***

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do Título de Doutor em Engenharia.

Área de Concentração: Sistemas Eletrônicos

Orientador: Prof. Dr. Sergio Takeo Kofuji

São Paulo
2006

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO OU PESQUISA, DESDE QUE CITADA A FONTE.

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, de abril de 2006.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Ferreira, Eveline Cruz Hora Gomes

Geração automática de metadados: uma contribuição para a Web semântica / E.C.H.G. Ferreira. -- ed.rev. -- São Paulo, 2006. 229 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Semântica 2.Biblioteca digital 3.Recuperação da informação 4.Descoberta de conhecimento I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

***“Se depender de mim,
nunca ficarei plenamente maduro,
nem nas idéias nem no estilo,
mas sempre verde,
incompleto,
experimental”***

*Gilberto Freire,
“Tempo morto e outros tempos”,
anotação de 1926*

Aos meus filhos
Rodrigo, Felipe, Matheus e Débora
e a meu esposo Augusto.

Agradecimentos

Agradeço primeiramente a Deus, pela oportunidade desta vida repleta de alegrias e realizações. Agradeço a Jesus, meu mestre e amigo, e aos meus anjos protetores, que trabalhando em silêncio guiaram meus pensamentos, iluminando meu caminho.

Ao Prof. Dr. Sergio Takeo Kofuji, por sua orientação e amizade, e pelos anos de convivência que muito contribuíram para o meu crescimento científico e intelectual.

À Fundação Joaquim Nabuco, pela oportunidade de realização do curso de Doutorado.

Aos colegas do LSI e a Marcelo Zenaide, pelas várias vezes que interromperam seus trabalhos para me ajudar, pelos momentos de descontração e pelo carinho.

Aos colegas da Fundação Joaquim Nabuco, pelo apoio, incentivo e compreensão.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

Resumo

Abstract

1 INTRODUÇÃO	18
1.1 OBJETIVOS DO TRABALHO	19
1.2 JUSTIFICATIVA	20
1.3 METODOLOGIA	22
1.3.1 Estudos e Pesquisas Bibliográficas	22
1.3.2 Definição da Arquitetura Geral do Modelo SIGAMeta	23
1.3.3 Execução do Modelo SIGAMeta	26
1.3.4 Métricas e Métodos de Avaliação	27
1.4 CONTRIBUIÇÕES	28
1.5 TRABALHOS RELACIONADOS	29
1.6 ORGANIZAÇÃO DO TRABALHO	34
2 AMBIENTES DIGITAIS E A WEB SEMÂNTICA	36
2.1 AMBIENTES DIGITAIS VOLTADOS À GESTÃO E DISSEMINAÇÃO DE INFORMAÇÃO	36
2.2 A WEB ATUAL E A WEB SEMÂNTICA	40
2.3 COMPONENTES FUNDAMENTAIS DA WEB SEMÂNTICA	42
2.3.1 O XML - eXtensible Markup Language	43
2.3.2 O RDF - Resource Description Framework	47
2.3.3 Ontologia	54
2.3.4 Metadados	57
2.4 DESAFIOS NA CONSTRUÇÃO E MANUTENÇÃO DE UM AMBIENTE DIGITAL SEMÂNTICO ..	57
3 METADADOS: APENAS DADOS SOBRE DADOS?	59
3.1 DEFINIÇÃO	59
3.2 USOS E FUNÇÕES	60
3.3 CATEGORIAS DOS METADADOS	62
3.4 CRIAÇÃO E GERENCIAMENTO DE METADADOS EM AMBIENTE DIGITAL	65
3.5 PADRONIZAÇÃO DE METADADOS NO CONTEXTO DA WEB	68
4 DESCOBERTA DE CONHECIMENTO EM TEXTO	72
4.1 RECUPERAÇÃO DE INFORMAÇÃO	73
4.1.1 Modelo Booleano	77

4.1.2 Modelo Vetorial	78
4.1.3 Modelo Booleano Estendido	84
4.2 RECUPERAÇÃO DE INFORMAÇÃO NA WEB	88
4.3 DESCOBERTA DE CONHECIMENTO EM TEXTO	91
4.4 CATEGORIZAÇÃO DE TEXTOS	95
4.5 A SUMARIZAÇÃO DE TEXTOS	98
4.5.1 A Sumarização Automática	100
5 SIGAMETA - UM MODELO DE GERAÇÃO AUTOMÁTICA DE METADADOS ...	108
5.1 CORPORA DE DOCUMENTOS TEXTUAIS A SEREM PROCESSADOS PELO MODELO SIGAMETA	108
5.1.1 Corpora de Treinamento em Categorias Específicas	108
5.1.2 Corpora para Geração dos Metadados	109
5.2 ELEMENTOS DE METADADOS GERADOS PELO MODELO SIGAMETA	110
5.3 FUNCIONAMENTO GERAL DO MODELO SIGAMETA E INTERFACES GRÁFICAS	112
5.3.1 Módulo de Configuração	113
5.3.2 Módulo de Extração de Padrões de Categorias	114
5.3.3 Módulo de Geração de Metadados	115
5.4 ARQUITETURA GERAL E DETALHADA DO MODELO SIGAMETA	116
5.4.1 Detalhamento da Fase de Treinamento	117
5.4.2 Detalhamento da Fase de Geração de Metadados	121
5.5 MÉTRICAS E MÉTODOS DE AVALIAÇÃO DO MODELO SIGAMETA	126
6 AVALIAÇÃO E ANÁLISE DOS RESULTADOS	130
6.1 RESULTADOS PRODUZIDOS PELO MÓDULO DE EXTRAÇÃO DE PADRÕES DE CATEGORIAS	130
6.2 RESULTADOS PRODUZIDOS PELO MÓDULO DE GERAÇÃO DE METADADOS	136
6.2.1 Avaliação do Metadado Subject	137
6.2.2 Avaliação do Metadado Description	151
6.2.3 Avaliação dos Documentos RDF/XML	159
7 CONCLUSÕES E TRABALHOS FUTUROS	167
7.1 CONCLUSÕES	169
7.2 LIMITAÇÕES	173
7.3 TRABALHOS FUTUROS	173
APÊNDICE A - DESAFIOS NA CONSTRUÇÃO E MANUTENÇÃO DE UM AMBIENTE DIGITAL SEMÂNTICO	175
A.1 DESAFIOS NA PROVISÃO DE ACESSOS EFETIVOS:	175
A.2 DESAFIOS NA CONSTRUÇÃO E MANUTENÇÃO DAS COLEÇÕES DIGITAIS:	180
A.3 DESAFIOS NA INTEROPERABILIDADE DE AMBIENTES DIGITAIS:	186
A.4 DESAFIOS EM DIREITOS AUTORAIS:	188
A.5 DESAFIOS NA FORMAÇÃO E SUSTENTAÇÃO DE RECURSOS HUMANOS:	190
A.6 DESAFIOS NA CAPTAÇÃO DE RECURSOS FINANCEIROS:	191
ANEXO A - DIAGRAMAS COMPORTAMENTAIS DO MODELO SIGAMETA	193
A.1 DIAGRAMA DE CASO DE USO	193
A.2 DIAGRAMA DE SEQUÊNCIA	194

ANEXO B - INFORMAÇÕES E RESULTADOS COMPLEMENTARES.....	197
B.1 LISTA DE <i>STOPWORDS</i>	197
B.2 VETORES DE CATEGORIAS	198
B.3 PALAVRAS E <i>STEMMINGS</i> RSLP E PORTER	204
B.4 TEXTOS ORIGINAIS	207
B.5 DOCUMENTOS RDF/XML	211
B.6 DOCUMENTO XSL	214
B.7 PÁGINA <i>WEB</i>	219
BIBLIOGRAFIA	221

Lista de Figuras

FIGURA 2.1 – O <i>ROADMAP</i> DA WEB SEMÂNTICA [SEMANTICWEB].....	41
FIGURA 2.2 – ARQUITETURA EM CAMADAS DA WEB SEMÂNTICA [BLH2001]...	43
FIGURA 2.3 – EXEMPLO DE UM DOCUMENTO EM HTML	43
FIGURA 2.4 - EXEMPLO DE UM DOCUMENTO EM XML	45
FIGURA 2.5 – ESTRUTURA LÓGICA INTERNA DE DOCUMENTOS XML	45
FIGURA 2.6 – EXEMPLO DE UM DTD PARA UMA AGENDA.....	46
FIGURA 2.7 – DECLARAÇÃO RDF EM FORMATO XML	49
FIGURA 2.8 – REPRESENTAÇÃO EM FORMATO DE TRIPLA PARA UM DOCUMENTO RDF/XML	50
FIGURA 2.9 – REPRESENTAÇÃO EM FORMATO DE GRAFO PARA UM DOCUMENTO RDF/XML	51
FIGURA 2.10 – GRAFO DE UMA HIERARQUIA DE CLASSES RDFS	52
FIGURA 2.11 – CÓDIGO RDFS	53
FIGURA 3.1 – O CICLO DE VIDA DE OBJETOS CONTIDOS EM UM SISTEMA DE INFORMAÇÃO DIGITAL [GILLILAND2000]	66
FIGURA 4.1 - COMPONENTES DE UM PROCESSO DE RECUPERAÇÃO DE INFORMAÇÃO	74
FIGURA 4.2 - MATRIZ DE VETORES E COMBINAÇÃO BOOLEANA DE TERMOS DE INDEXAÇÃO.....	77
FIGURA 4.3 – DISTÂNCIA EUCLIDIANA	78
FIGURA 4.4 – REPRESENTAÇÃO DE UMA ESPECIFICAÇÃO DE CONSULTA E DOIS DOCUMENTOS EM UM ESPAÇO VETORIAL BIDIMENSIONAL	79
FIGURA 4.5 – MATRIZ DE PESOS DE <i>CORPUS</i> CONTENDO <i>I</i> DOCUMENTOS E <i>T</i> TERMOS DE INDEXAÇÃO	80
FIGURA 4.6 - DISTÂNCIA EUCLIDIANA MÁXIMA EM ESPAÇO BIDIMENSIONAL	85

FIGURE 4.7 - DISTÂNCIA EUCLIDIANA EM ESPAÇO BIDIMENSIONAL (OPERADOR OR).....	86
FIGURE 4.8 - DISTÂNCIA EUCLIDIANA EM ESPAÇO BIDIMENSIONAL (OPERADOR AND)	86
FIGURA 4.9 – MEDIDA DE SIMILARIDADE NORMALIZADA (OPERADORES OR E AND)	87
FIGURA 4.10 – MEDIDA DE SIMILARIDADE NORMALIZADA PARA <i>M</i> TERMOS (OR E AND)	87
FIGURA 4.11 – MEDIDA DE SIMILARIDADE COM PESOS DA ESPECIFICAÇÃO DE CONSULTA (OR E AND).....	88
FIGURA 4.12 – FLUXOGRAMA DO ALGORITMO <i>STEMMER PORTUGUESE</i> (RSLP)	94
FIGURA 4.13 – COMPONENTES DE UM PROCESSO AUTOMÁTICO DE CATEGORIZAÇÃO ASSISTIDA DE TEXTOS	97
FIGURA 5.1 – INTERFACE GRÁFICA DO MÓDULO DE CONFIGURAÇÃO DO MODELO	113
FIGURA 5.2 – INTERFACE GRÁFICA DO MÓDULO DE EXTRAÇÃO DE PADRÕES DE CATEGORIAS.....	115
FIGURA 5.3 – INTERFACE GRÁFICA DO MÓDULO DE GERAÇÃO DE METADADOS	116
FIGURA 5.4 – DIAGRAMA GERAL DO MODELO	117
FIGURA 5.5 – AMPLIAÇÃO DO MÓDULO DE EXTRAÇÃO DE PADRÕES DAS CATEGORIAS	118
FIGURA 5.6 – AMPLIAÇÃO DO MÓDULO DE GERAÇÃO DE METADADOS	122
FIGURA 5.7 – INTERFACE GRÁFICA DO APLICATIVO DE RECUPERAÇÃO DE INFORMAÇÃO	127
FIGURA 6.1 - TAXA DE ACERTO NA CATEGORIZAÇÃO DOS <i>CORPORA</i> PARA GERAÇÃO DOS METADADOS – RSLP	147
FIGURA 6.2 - TAXA DE ACERTO NA CATEGORIZAÇÃO DOS <i>CORPORA</i> PARA GERAÇÃO DOS METADADOS - PORTER	147
FIGURA 6.3 - TAXA DE ACERTO NA CATEGORIZAÇÃO DOS <i>CORPORA</i> PARA GERAÇÃO DOS METADADOS (PORTER X RSLP).....	148

FIGURA 6.4 – QUALIDADE DOS SUMÁRIOS EXTRATIVOS (RSLP)	156
FIGURA 6.5 – QUALIDADE DOS SUMÁRIOS EXTRATIVOS (PORTER)	156
FIGURA 6.6 – QUALIDADE DOS SUMÁRIOS EXTRATIVOS PARA TAXA DE COMPRESSÃO DE 70% (RSLP)	157
FIGURA 6.7 – QUALIDADE DOS SUMÁRIOS EXTRATIVOS PARA TAXA DE COMPRESSÃO DE 70% (PORTER)	157
FIGURA 6.8 – MEDIDAS DE PRECISÃO OBTIDAS EM BUSCAS A DOCUMENTOS RDF/XML E DOCUMENTOS TEXTO	165
FIGURA 6.9 – MEDIDAS DE REVOCACÃO OBTIDAS EM BUSCAS A DOCUMENTOS RDF/XML E DOCUMENTOS TEXTO	166
FIGURA A.1 – DIAGRAMA DE CASO DE USO	194
FIGURA A.2 – DIAGRAMA DE SEQÜÊNCIA DA OPERAÇÃO <i>PROCESSAR</i> <i>CATEGORIA</i>	195
FIGURA A.3 – DIAGRAMA DE SEQÜÊNCIA DA OPERAÇÃO <i>PROCESSAR</i> <i>DOCUMENTO</i>	196
FIGURA B.1 – EXIBIÇÃO DAS INFORMAÇÕES DE UM DOCUMENTO RDF/XML, NA <i>WEB</i>	220

Lista de Tabelas

TABELA 1.1 – ELEMENTOS DE METADADOS DC GERADOS PELO MODELO DE FORMA AUTOMÁTICA	23
TABELA 1.2 – ELEMENTOS DE METADADOS DC GERADOS PELO MODELO DE FORMA SEMI-AUTOMÁTICA	23
TABELA 1.3 – TABELA COMPARATIVA DOS TRABALHOS RELACIONADOS E TRABALHO PROPOSTO	33
TABELA 3.1 – CATEGORIAS DE METADADOS QUE REFLETEM PERSPECTIVAS DE GERENCIAMENTO E RECUPERAÇÃO DE RECURSOS EM AMBIENTE DIGITAL, E SUAS FUNÇÕES.....	63
TABELA 3.2 – ATRIBUTOS E CARACTERÍSTICAS DE METADADOS.....	63
TABELA 3.3 – ELEMENTOS DE METADADOS DO <i>DUBLIN CORE</i> (HTTP://PURL.PT/201/1/#DCT1)	70
TABELA 5.1 – ELEMENTOS DE METADADOS DC GERADOS PELO MODELO DE FORMA AUTOMÁTICA	111
TABELA 5.2 – ELEMENTOS DE METADADOS DC GERADOS PELO MODELO DE FORMA SEMI-AUTOMÁTICA	111
TABELA 6.1 – <i>CORPORA</i> DE TREINAMENTO EM CATEGORIAS ESPECÍFICAS ...	131
TABELA 6.2 – CARACTERÍSTICAS DOS VETORES DE CATEGORIAS.....	132
TABELA 6.3 – SEIS PRIMEIROS <i>STEMMINGS</i> DA CATEGORIA CPOLI - RSLP.....	133
TABELA 6.4 – SEIS PRIMEIROS <i>STEMMINGS</i> DA CATEGORIA ECONO - RSLP	133
TABELA 6.5 – SEIS PRIMEIROS <i>STEMMINGS</i> DA CATEGORIA ESTPOP - RSLP....	133
TABELA 6.6 – SEIS PRIMEIROS <i>STEMMINGS</i> DA CATEGORIA ANTROP - RSLP ..	134
TABELA 6.7 – PERCENTUAL DE REDUÇÃO DE PALAVRAS A SEU <i>STEMMING</i> (PORTER X RSLP).....	134
TABELA 6.8 – EXEMPLOS DE <i>STEMMINGS</i> RSLP E PORTER.....	135
TABELA 6.9 – <i>CORPORA</i> PARA GERAÇÃO DOS METADADOS	137
TABELA 6.10– PADRÃO DE ERROS DA CATEGORIZAÇÃO PARA DMS = 0,0 E 0,01	138

TABELA 6.11 – PADRÃO DE ERROS DA CATEGORIZAÇÃO PARA DMS = 0,02....	139
TABELA 6.12 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (RSLP, DMS = 0,0)	141
TABELA 6.13 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (PORTER, DMS = 0,0)	141
TABELA 6.14 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (RSLP, DMS = 0,01)	143
TABELA 6.15 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (PORTER, DMS = 0,01)	143
TABELA 6.16 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (RSLP, DMS = 0,02)	145
TABELA 6.17 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE DOCUMENTOS (PORTER, DMS = 0,02)	145
TABELA 6.18 – <i>STEMMINGS</i> E PESOS PRODUZIDOS PARA UM TEXTO EXEMPLO	151
TABELA 6.19 – CARACTERÍSTICAS DOS TEXTOS PARA GERAÇÃO DE SUMÁRIOS EXTRATIVOS	152
TABELA 6.20 – CRITÉRIOS PARA ATRIBUIÇÃO DE QUALIDADE AOS SUMÁRIOS	152
TABELA 6.21 – AVALIAÇÃO HUMANA DOS SUMÁRIOS EXTRATIVOS	153
TABELA 6.22 – QUALIDADE DOS SUMÁRIOS EXTRATIVOS	155
TABELA 6.23 – TAXAS DE ACERTO NA CATEGORIZAÇÃO DE SUMÁRIOS EXTRATIVOS (PORTER, DMS= 0,0; TAXA DE COMPRESSÃO DE 70%).....	158
TABELA 6.24 – <i>CORPORA</i> DE TEXTOS PARA GERAÇÃO DE DOCUMENTOS RDF/XML.....	161
TABELA 6.25 – <i>CORPORA</i> DE DOCUMENTOS RDF/XML.....	162
TABELA 6.26 – RESULTADOS DAS BUSCAS COM BASE NOS DOCUMENTOS RDF/XML E DOCUMENTOS TEXTO	163
TABELA 6.27 – RESULTADOS DAS BUSCAS COM BASE NOS DOCUMENTOS RDF/XML E DOCUMENTOS TEXTO	164
TABELA B.1 – LISTA DE <i>STOPWORDS</i>	197

TABELA B.2 – VETOR DE CATEGORIA CPOLI.....	199
TABELA B.3 – VETOR DE CATEGORIA ECONO	200
TABELA B.4 – VETOR DE CATEGORIA ESTPOP	202
TABELA B.5 – VETOR DE CATEGORIA ANTROP	203
TABELA B.6 – PALAVRAS E <i>STEMMINGS</i> DA CATEGORIA ANTROP	205

Lista de Abreviaturas

AACR	<i>Anglo-American Cataloging Rules</i>
AACR2	<i>Anglo-American Cataloging Rules (2ª edição)</i>
ABEP	<i>Associação Brasileira de Estudos Populacionais</i>
ADL	<i>Advanced Distributed Learning</i>
ANTROP	<i>Antropologia Cultural</i>
CERN	<i>Centre Européen de Recherche Nucléaire</i>
CPDOC	<i>Centro de Pesquisa e Documentação de História Contemporânea do Brasil</i>
CPOLI	<i>Ciências Políticas</i>
CSS	<i>Cascading Style Sheets</i>
DAML	<i>Sistema de Informação Geográfica</i>
DC	<i>Dublin Core</i>
DCMES	<i>Dublin Core Metadata Element Set</i>
DCMI	<i>Dublin Core Metadata Initiative</i>
DDC	<i>Dewey Decimal Classification</i>
D_i	<i>Documento i</i>
DMS	<i>Distância Máxima de Similaridade</i>
DOI	<i>Digital Object Identifier</i>
DSS	<i>Sistemas de Suporte à Decisão</i>
DTD	<i>Document Type Definition</i>
ECONO	<i>Economia</i>
EI	<i>Extração de Informação</i>
ESTPOP	<i>Estudos Populacionais</i>
FGDC	<i>Content Standard for Digital Geospatial Metadata</i>
FJN	<i>Fundação Joaquim Nabuco</i>
FUNDAJ	<i>Fundação Joaquim Nabuco</i>
GILS	<i>Global Information Locator Service</i>
HTML	<i>Hiper Text Markup Language</i>
HTTP	<i>Hiper Text Transfer Protocol</i>
IA	<i>Inteligência Artificial</i>
idf	<i>Inverse Document Frequency</i>
ISBN	<i>International Standard Book Number</i>
isf	<i>Inverse Sentence frequency</i>
ISO	<i>International Standards Organization</i>
KB	<i>KiloBite</i>
KDD	<i>Knowledge Discovery in Database</i>
KDT	<i>Knowledge Discovery from Text</i>
LTSC	<i>Learning Technology Standards Committee</i>
MARC	<i>Machine Readable Cataloging</i>
MB	<i>MegaByte</i>
N	<i>Número de documentos em um corpus</i>
n_k	<i>Número de documentos que possui o termo T_k</i>
OIL	<i>Ontology Interchange Language</i>
OMG	<i>Object Management Group</i>

PDA	<i>Personal Digital Assistants</i>
PDF	<i>Portable Document Format</i>
PLN	<i>Processamento de Linguagem Natural</i>
P_{ik}	<i>Número de palavras que compõe o termo T_k</i>
Q	<i>Especificação de consulta</i>
RDF	<i>Resource Description Framework</i>
RDF(S)	<i>RDF + RDFS</i>
RDFS	<i>RDF Schema</i>
RI	<i>Recuperação de Informação</i>
RSLP	<i>Removedor de Sufixos para a Língua Portuguesa</i>
SA	<i>Sumarização Automática</i>
SGML	<i>Standard Generalized Markup Language</i>
SHOE	<i>Simple HTML Ontology Extensions</i>
SIG	<i>Sistema de Informação Geográfica</i>
SIGAMeta	<i>Sistema de Geração Automática de Metadados</i>
$Sim(D,Q)$	<i>Similaridade entre D e Q</i>
Sim-min	<i>Similaridade mínima</i>
tf	<i>Term Frequency – Frequência Absoluta</i>
TIC	<i>Tecnologias de Informação e de Comunicação</i>
T_k	<i>Termo k de um documento</i>
txt	<i>Formato texto</i>
UML	<i>Unified Modeling Language</i>
URI	<i>Uniform Resource Identifier</i>
URL	<i>Uniform Resource Locator</i>
URN	<i>Universal Resource Name/Number</i>
VHS	<i>Video Home System</i>
W3C	<i>World Wide Web Consortium</i>
W_{ik}	<i>Peso do termo T_k no documento D_i</i>
WWLib	<i>Wolverhampton Web Library</i>
WWW	<i>World Wide Web</i>
XML	<i>eXtensible Markup Language</i>
XOL	<i>Ontology Exchange Language</i>
XSL	<i>eXtensible Stylesheet Language</i>

Resumo

FERREIRA, E. C. H. G. **Geração Automática de Metadados: uma Contribuição para a Web Semântica**. 2006. 229 p. Tese (Doutorado). Escola Politécnica da Universidade de São Paulo, 2006.

Esta Tese oferece uma contribuição na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*, definindo um *Modelo de geração automática de metadados baseado em contexto*, a partir de documentos textuais na língua portuguesa, em formato não estruturado (txt). Um conjunto teórico amplo de assuntos ligados à criação de ambientes digitais semântico também é apresentado.

Conforme recomendado em *SemanticWeb.org*, os documentos textuais aqui estudados foram automaticamente convertidos em páginas *Web* anotadas semanticamente, utilizando o *Dublin Core* como padrão para definição dos elementos de metadados, e o padrão RDF/XML para representação dos documentos e descrição dos elementos de metadados.

Dentre os quinze elementos de metadados *Dublin Core*, nove foram gerados automaticamente pelo Modelo, e seis foram gerados de forma semi-automática. Os metadados *Description* e *Subject* foram os que necessitaram de algoritmos mais complexos, sendo obtidos através de técnicas estatísticas, de mineração de textos e de processamento de linguagem natural.

A finalidade principal da avaliação do Modelo foi verificar o comportamento dos documentos convertidos para o formato RDF/XML, quando estes foram submetidos a um processo de recuperação de informação. Os elementos de metadados *Description* e *Subject* foram exaustivamente avaliados, uma vez que estes são os principais responsáveis por apreender a semântica de documentos textuais.

A diversidade de contextos, a complexidade dos problemas relativos à língua portuguesa, e os novos conceitos introduzidos pelos padrões e tecnologias da Web Semântica, foram alguns dos fortes desafios enfrentados na construção do Modelo aqui proposto.

Apesar de se ter utilizado técnicas não muito novas para a exploração dos conteúdos dos documentos, não se pode ignorar que os elementos inovadores introduzidos pela Web Semântica ofereceram avanços que possibilitaram a obtenção de resultados importantes nesta Tese.

Como demonstrado aqui, a junção dessas técnicas com os padrões e tecnologias recomendados pela Web Semântica pode minimizar um dos maiores problemas da *Web* atual, e uma das fortes razões para a implementação da Web Semântica: a tendência dos mecanismos de busca de inundarem os usuários com resultados irrelevantes, por não levarem em consideração o contexto específico desejado pelo usuário.

Dessa forma, é importante que se dê continuidade aos estudos e pesquisas em todas as áreas relacionadas à implementação da Web Semântica, dando abertura para que sistemas de informação mais funcionais sejam projetados.

Abstract

FERREIRA, E. C. H. G. **Automatic Metadata Generation: a Contribution to the Semantic Web.** 2006. 229 p. Thesis (Doctoral). Escola Politécnica da Universidade de São Paulo, 2006.

This Thesis offers a contribution to the Semantic Web area, in the scope of the *representation and indexing of documents*, defining an *Automatic metadata generation model based on context*, starting from textual documents not structured in the Portuguese language. A wide theoretical set of subjects related to the creation of semantic digital environments is also presented.

As recommended in *SemanticWeb.org*, the textual documents studied here were automatically converted to *Web* pages written in semantic format, using *Dublin Core* as standard for definition of metadata elements, and the standard RDF/XML for representation of documents and description of the metadata elements.

Among the fifteen *Dublin Core* metadata elements, nine were automatically generated by the Model, and six were generated in a semiautomatic manner. The metadata *Description* and *Subject* were the ones that required more complex algorithms, being obtained through statistical techniques, text mining techniques and natural language processing.

The main purpose of the evaluation of the Model was to verify the behavior of the documents converted to the format RDF/XML, when these were submitted to an information retrieval process. The metadata elements *Description* and *Subject* were exhaustively evaluated, since these are the main ones responsible for learning the semantics of textual documents.

The diversity of contexts, the complexity of the problems related to the Portuguese language, and the new concepts introduced by the standards and technologies of the Semantic Web, were some of the great challenges faced in the construction of the Model here proposed.

In spite of having used techniques which are not very new for the exploration and exploitation of the contents of the documents, we cannot ignore that the innovative elements introduced by the Web Semantic have offered improvements that made possible the obtention of important results in this Thesis.

As demonstrated here, the joining of those techniques with the standards and technologies recommended by the Semantic Web can minimize one of the largest problems of the current Web, and one of the strong reasons for the implementation of the Semantic Web: the tendency of the search mechanisms to flood the users with irrelevant results, because they do not take into account the specific context desired by the user. Therefore, it is important that the studies and research be continued in all of the areas related to the Semantic Web's implementation, opening the door for more functional systems of information to be designed.

1

Introdução

Avanços recentes em tecnologias de informação e de comunicação (TIC), juntamente com os custos cada vez menores de equipamentos e sistemas informatizados, vêm apresentando oportunidades interessantes para criação, conversão, comercialização, distribuição e uso de conteúdos digitais, substituindo a mídia impressa ou analógica em uma grande variedade de aplicações. Com a tecnologia digital, as informações passaram a ter uma maior longevidade, disponibilidade e aperfeiçoamento em suas buscas e descobertas, possibilitando o acesso ao conteúdo procurado em qualquer momento e onde quer que se encontre o usuário [Gilliland2000, Borg1999, Bullock1999]. Neste contexto, entidades possuidoras de informação como bibliotecas, museus, arquivos e instituições de ensino e pesquisa em geral, vêm tendo a oportunidade de reforçar o cumprimento de suas responsabilidades na geração, disseminação, uso e preservação da informação, de uma maneira inovadora, efetiva e oportuna.

A forma mais usual de acesso à informação digital tem sido através da *Web*. Surgida no início dos anos 90, a *Web* buscava oferecer interfaces mais amigáveis e intuitivas para a organização e acesso ao crescente repositório de documentos que se tornava a Internet. Entretanto, o enorme crescimento desta rede, além da ampliação das possibilidades de sua utilização, tem apontado a necessidade de uma nova filosofia de estruturação e acesso, além da ampliação de sua infra-estrutura tecnológica de comunicação [SA2004].

Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e recuperação de informações para usuários humanos, a *Web* foi implementada de forma descentralizada e quase anárquica. Cresceu exponencialmente e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando se quer recuperar a informação necessária. Não há estratégia satisfatória para a indexação dos documentos, e a recuperação das informações, possível por meio dos “motores de busca”, é baseada principalmente em palavras-chave contidas nos textos, o que é muito pouco eficaz quando

não se tem uma estrutura de dados que se explicita o contexto das informações. Além disso, os motores de busca só conseguem dar uma cobertura limitada e pequena das informações contidas na *Web*, de forma que a grande maioria de seus documentos permanece encoberta.

A dificuldade de se explicitar o contexto dos documentos disponíveis na *Web* tem exigido do usuário uma análise mais profunda de seus resultados. Além disso, a ênfase das tecnologias e linguagens atualmente utilizadas nas páginas *Web* (como o HTML) focaliza os aspectos de apresentação dos dados, sem expor o seu significado, de forma que a informação se torna pobremente descrita e pouco passível de ser consumida por máquinas e seres humanos. É neste contexto que surge a proposta da *Web Semântica*.

A Web Semântica (W3C) vem se apresentando como solução para organizar as informações existentes na Web. Segundo Berners-Lee [BLH2001], a Web Semântica será uma extensão da *Web* atual, porém apresentará estrutura que possibilitará a compreensão e o gerenciamento dos conteúdos armazenados na *Web*, tanto por pessoas como por programas, independente da forma em que estes se apresentem (texto, som, imagem e gráficos). Isso tudo será possível a partir da valoração semântica dos conteúdos dos documentos, e através de agentes que são programas coletores de conteúdos advindos de fontes diversas, capazes de processar as informações e permutar automaticamente resultados com outros programas.

1.1 Objetivos do Trabalho

Esta Tese tem como objetivo fornecer uma contribuição na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*.

A *Representação de documentos* objetiva descrever ou identificar cada documento através de estruturas de dados conceituais (metadados) extraídas de seus conteúdos, para efeito de economia de armazenamento, agilidade e efetividade nas buscas e descobertas.

Mais especificamente, o trabalho define um *Modelo de geração automática de metadados baseado em contexto, a partir de documentos textuais na língua portuguesa, em formato não estruturado (txt)*. Estes metadados visam o *gerenciamento e recuperação de recursos de informação na Web*.

No Modelo proposto (aqui denominado SIGAMeta), conforme recomendado em *SemanticWeb.org* [SemanticWeb], os documentos textuais são automaticamente convertidos em páginas *Web* anotadas semanticamente, utilizando padrões e tecnologias que

proporcionam automação, compartilhamento, integração e reuso de informação para várias aplicações. Estes padrões darão abertura para que sistemas de informação mais funcionais sejam projetados, como por exemplo, para recuperação e acesso à informação, gerenciamento, preservação e proteção de direitos autorais.

1.2 Justificativa

Na criação de ambientes digitais efetivos, tecnologias de informação e de comunicação têm oferecido oportunidades inestimáveis de avanço, especialmente na área de Recuperação de Informação (RI).

A área de RI cresceu muito em importância com a introdução da *Web*. Apesar de seu sucesso, a *Web* introduziu uma série de problemas por si mesma. O principal deles é a ausência de mecanismos que proporcionem buscas e descobertas dentro de um contexto específico desejado, exigindo que usuários analisem cada uma das informações individualmente, e descartem aquelas (geralmente maioria) que não atendem os seus interesses (baixa precisão) [LR2001, LWO2000].

A tendência de tais mecanismos para inundar os usuários com resultados irrelevantes tem mostrado aos pesquisadores da área a necessidade de padrões que representem e disponibilizem adequadamente os documentos, bem como de sistemas de indexação eficazes. Esses elementos minimizariam problemas de ambigüidade de informações em buscas e descobertas, não apenas para usuários humanos, mas também para programas de computador. É neste âmbito que a proposta da Web Semântica e suas tecnologias adjacentes aparecem como solução promissora.

A Web Semântica é um esforço colaborativo liderado pela *World Wide Web Consortium* [W3C-SW], com participação de um grande número de pesquisadores e parceiros industriais. Uma de suas implementações está baseada no *Resource Description Framework* (RDF) [W3C-RDF], que usa notação XML (*eXtensible Markup Language*) [W3C-XML] como sintaxe de codificação para descrição de metadados (combinação conhecida como RDF/XML).

Na realização da Web Semântica serão necessárias padronizações de tecnologias, de linguagens e de metadados, de forma que todos os usuários da *Web* possam obedecer a

regras comuns sobre como armazenar e descrever a informação. Dessa forma, informações poderão ser “consumidas” por outros usuários humanos ou não, de maneira automática e não ambígua. Conclui-se então que não apenas o material a ser disponibilizado pela primeira vez na *Web* deverá estar de acordo com esses novos padrões, como também será necessária a conversão do material já existente na *Web*.

Para ambos os casos, processos automatizados e/ou semi-automatizados de estruturação, conversão de padrões e interpretação semântica do conteúdo serão necessários, livrando o usuário do custo de manipulação de enormes volumes de dados, bem como evitando erros comuns de digitação e interpretação de conteúdo [Greenberg2003].

A Web Semântica não se trata de uma revolução, mas sim de uma evolução da *Web* atual. Trata-se principalmente da adoção de padrões de metadados e de compartilhamento destes padrões, de forma que se possa melhor utilizar o vasto repositório de informações disponível da *Web* de maneira mais produtiva, ágil e significativa.

Metadados, tendo o potencial de registrar conteúdo, contexto e estrutura de recursos de informação, têm oferecido meios efetivos de viabilizar não apenas atividades de recuperação de informação, mas também diversas outras, como gerenciamento, preservação, disseminação e direitos autorais [Tannenb2001]. Entretanto, metadados, em sua grande maioria, vêm sendo gerados manualmente, o que tem trazido sérias dificuldades no desenvolvimento de soluções que explorem suas potencialidades. A geração manual de metadados é dispendiosa, além de passível de erros de digitação e emprego de vocabulário fora do padrão adotado [AMEGA2005].

A necessidade de técnicas automatizadas para criação de metadados, e/ou que facilitem a sua geração manual, tem sido reconhecida por várias comunidades envolvidas na produção, gerenciamento, distribuição, uso e preservação de recursos de informação, e estudos ainda se encontram em fase inicial de pesquisa [AMEGA2005, BARBOSA2003, JJBW1998, PTC1999, SMZ2003, SZR2000, YSLH2003], principalmente no âmbito de documentos da língua portuguesa.

Segundo a referência [Ethnologue], a língua portuguesa é o quinto maior idioma em número de oradores nativos. Há cerca de 177 milhões de oradores nativos e outros 15 milhões de pessoas que a utilizam como um segundo idioma. É falado em quatro continentes: Europa (Portugal, Madeira, o Açores), América do Sul (Brasil), África (Angola,

Moçambique, Ilhas de Cabo Verde, Guiné - Bissau) e Ásia (Goa, Macau, Timor Oriental). Em contraste, de acordo com a *Internet Systems Consortium* [ISC], menos que 1% de todos os *Web Hosts* estão em português. Além disso, apenas uma pequena porcentagem da população que fala a língua portuguesa é competente em inglês, e a grande maioria (inclusive graduados) não é capaz de apresentar uma questão em inglês para uma máquina de pesquisa como *Google*, *Altavista*, *Lycos* etc. Como consequência, fontes de informação para os 8 milhões de oradores da língua portuguesa que têm acesso à Internet estão extremamente limitadas, comparadas com a imensa quantidade de informação disponível à população que fala inglês [Orengo2004].

Neste sentido, o trabalho aqui apresentado vem incentivar os estudos e pesquisas na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*, propondo um *Modelo de geração automática de metadados baseado em contexto, a partir de documentos textuais na língua portuguesa, em formato não estruturado (txt)*.

1.3 Metodologia

Com o objetivo de estudar a geração automática de metadados que visam gerenciamento e recuperação de informação baseada em contexto, o trabalho dividiu-se em quatro etapas cujas atividades não necessariamente se encontram em ordem de execução.

1.3.1 Estudos e Pesquisas Bibliográficas

Esta primeira etapa teve o objetivo de:

- Averiguar o estado da arte de assuntos gerais sobre ambientes digitais semânticos, e mais especificamente sobre a Web Semântica e suas tecnologias inerentes (XML, RDF, Metadados e Ontologias);
- Aprofundar estudos sobre técnicas de Inteligência Artificial (IA) para descoberta de conhecimento (mais especificamente na área de Mineração de texto – *Text Mining* e Processamento de Linguagem Natural - PLN), bem como sobre tecnologias de Recuperação de Informação (RI) na *Web* e de Extração de Informação (EI);
- Averiguar as possíveis inovações em modelos já existentes de geração automática de metadados, bem como nas formas de categorização e sumarização de documentos

textuais, levando em consideração os novos conceitos introduzidos pela Web Semântica e por suas tecnologias.

1.3.2 Definição da Arquitetura Geral do Modelo SIGAMeta

O Modelo proposto contou com os seguintes padrões:

- Padrão para definição dos elementos de metadados: *Dublin Core Metadata Element Set*;
- Padrão para descrição dos elementos de metadados: RDF/XML;
- Padrão para representação dos documentos *Web*: RDF/XML.

a) Elementos de Metadados

Os elementos de metadados *Dublin Core* [DublinCore, SVM2000] formam um conjunto de quinze elementos. Destes, nove foram gerados automaticamente pelo Modelo proposto (Tabela 1.1), e seis foram gerados de forma semi-automática (Tabela 1.2).

Tabela 1.1 – Elementos de Metadados DC Gerados pelo Modelo de forma Automática

Elemento	Conteúdo gerado	Tipo/Função
Subject	Palavras-chave e categoria do documento	Descritivo
Description	Extrato	Descritivo
Publisher	FJN	Administrativo
Date	(AAAA-MM-DD) - Data de criação do documento digital	Técnico
ObjectType	text	Técnico
Format	text/xml	Técnico
Identifier	Categoria+seqüencial	Descritivo
Language	pt-br	Descritivo
Rights	FJN	Administrativo

Tabela 1.2 – Elementos de Metadados DC Gerados pelo Modelo de forma Semi-automática

Elemento	Conteúdo gerado	Tipo/Função
Title	“”	Descritivo
Author	“”	Descritivo
Contributor	“”	Descritivo
Relation	“”	Descritivo
Source	“”	Descritivo
Coverage	“”	Descritivo

A definição dos conteúdos dos metadados gerados automaticamente se baseou nas recomendações do *Dublin Core*, como será apresentado no Capítulo 3.

A escolha dos elementos do DC para geração automática foi baseada em três fatores: 1) importância na retenção de contexto (semântica); 2) possibilidade de extração do elemento a partir do conteúdo do documento; 3) possibilidade de extração do elemento a partir de parâmetros fornecidos pelo usuário ou pelo sistema operacional.

Os elementos de metadados DC *Description* e *Subject*, principais responsáveis em apreender a semântica de documentos textuais, foram obtidos através da análise do conteúdo dos documentos utilizando-se técnicas estatísticas, de mineração de textos e de processamento de linguagem natural. Os elementos restantes da Tabela 1.1, apesar de ausentes no conteúdo do documento, foram fornecidos pelo sistema operacional ou a partir de parâmetros do usuário, sendo então replicados automaticamente para todos os documentos processados.

Para os elementos de metadados *Title*, *Author*, *Contributor*, *Relation*, *Source* e *Coverage* (Tabela 1.2), o Modelo propôs a geração semi-automática, fornecendo a sua codificação em RDF/XML (com conteúdo vazio) para serem posteriormente preenchida de forma manual. Os elementos *Title* e *Author*, apesar de presentes no conteúdo de documentos, não oferecem pistas seguras para a sua correta identificação automática, quando se tratando de documentos não estruturados. Este procedimento semi-automático vem proporcionar a administradores de ambientes digitais uma maior flexibilidade na complementação de seus documentos.

b) Seleção dos Documentos Textuais

Para a realização dos experimentos, o Modelo contou com dois *Corpora*: *Corpora* de treinamento em categorias específicas e *Corpora* para geração dos metadados.

Corpora é o plural da palavra *corpus*. A definição clássica de *corpus* (*lato sensu*) é, segundo o Dicionário Aurélio, “um conjunto de documentos, dados e informações sobre determinada matéria”. Do ponto de vista da lingüística/semiologia, ainda na mesma obra, o termo é definido como “um conjunto finito de materiais significantes, constituído com vistas à análise semiológica”. Contudo, o termo já adquiriu uma nova definição, e é esta que será

utilizada nesta Tese daqui por diante (*corpus, stricto sensu*): “um conjunto de textos em formato eletrônico”.

- *Corpora* de treinamento em categorias específicas

Os *corpora de treinamento* foram compostos por 208 documentos textuais científicos não estruturados (formato txt), na língua portuguesa, pertencentes às categorias Ciências Políticas (61 textos), Economia (31 textos), Estudos Populacionais (35 textos) e Antropologia Cultural (81 textos).

Estes documentos foram recolhidos respectivamente das seguintes fontes: Centro de Pesquisa e Documentação de História Contemporânea do Brasil – CPDOC, da Fundação Getúlio Vargas (<http://www.cpdoc.fgv.br/comum/htm/index.htm>); Biblioteca Virtual de Economia do Ministério da Ciência e Tecnologia (<http://www.prossiga.br/nuca-ie-ufrj/economia/>); Associação Brasileira de Estudos Populacionais – ABEP (<http://www.abep.org.br/usuario/GerenciaNavegacao.php>); Folclore Brasileiro – Lendas, Crenças, Danças, Brincadeiras Infantis e Simpatias (<http://www.ifolclore.com.br/index.htm>).

- *Corpora* para geração dos metadados

Os *corpora para geração dos metadados*, igualmente aos *corpora* de treinamento, foram compostos por 119 documentos textuais científicos não estruturados (formato txt), na língua portuguesa, pertencentes às categorias Ciências Políticas (27 textos), Economia (21 textos), Estudos Populacionais (20 textos) e Antropologia Cultural (26 textos). Foram também considerados documentos pertencentes a outras categorias, objetivando verificar o comportamento do Modelo diante de categorias não previstas (25 textos). Todos esses documentos foram recolhidos do *site* da Fundação Joaquim Nabuco (www.fundaj.gov.br), nas opções *Textos on-line* e *Pesquisa Escolar*.

c) Implementação, análise de algoritmos e ajustes do Modelo

A linguagem de programação escolhida para a implementação do Modelo foi o Java, produzido pela *Sun Microsystems, Inc* (<http://java.sun.com/>). O Java é ao mesmo tempo um ambiente e uma linguagem de programação. Hoje é a linguagem que mais se fortalece com a Internet, pois é simples, orientada a objetos, distribuída, interpretada, robusta, segura,

independente de arquitetura, portátil, de alto desempenho, multitarefa e dinâmica. Todas estas características contribuíram positivamente não apenas na implementação do Modelo, como também na obtenção de um sistema robusto e eficiente.

Foram definidos e/ou adotados algoritmos e métodos para obtenção de cada um dos metadados, para converter os documentos textuais em páginas *Web* anotadas semanticamente e para auxiliar na avaliação dos metadados produzidos:

- Análise e adaptação de algoritmos já existentes para radicalização de palavras (*stemmer*) da língua portuguesa (Porter [Porter1980] e RSLP [OH2001]). *Stemmer* é uma técnica de padronização de vocabulário que busca identificar termos similares e reduzi-los a um único radical, eliminando suas diferenças morfológicas ou lexicais (flexões e variações da palavra);
- Definição de uma estratégia de atribuição de categorias a documentos: Categorização por diferença de Similaridade;
- Implementação de algoritmos para geração dos metadados utilizando técnicas estatísticas, mineração de texto e PLN;
- Implementação de algoritmo para geração dos documentos *Web* anotados semanticamente;
- Implementação de interfaces gráficas para configuração do sistema, definição de parâmetros de usuários, manutenção de bases de dados, além da execução do Modelo de forma integrada;
- Implementação de aplicativo de RI para avaliação de desempenho em buscas com base em *corpora* RDF/XML gerados pelo Modelo, e *corpora* em formato txt;
- Desenvolvimento de documento XSL para exibição dos dados na *Web*;
- Testes e ajustes do Modelo.

1.3.3 Execução do Modelo SIGAMeta

O Modelo foi concebido para ser executado em duas fases: Fase de treinamento e Fase de geração dos metadados.

A Fase de treinamento teve o objetivo de extrair, a partir de cada *corpus* dos *corpora* de treinamento, os elementos que representaram cada categoria (Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural).

A Fase de geração dos metadados executou os algoritmos que extraíram, para cada documento, cada um dos nove elementos de metadados do DC gerados automaticamente, a partir dos *corpora* para geração dos metadados e dos parâmetros de usuários. Os vetores de categorias gerados na primeira fase foram utilizados unicamente para a geração do metadado *Subject* (Categorização por Diferença de Similaridade).

Uma vez extraídos os metadados, o Modelo converteu cada documento para o formato RDF/XML, complementando-os com os metadados a serem gerados de forma semi-automática.

1.3.4 Métricas e Métodos de Avaliação

Uma vez que o objetivo desta Tese é “fornecer uma contribuição na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*”, a finalidade principal da avaliação do Modelo foi verificar, através das métricas de precisão e revocação, o comportamento dos documentos convertidos para o formato RDF/XML, quando estes foram submetidos a um processo de recuperação de informação.

O metadado *Description* foi obtido através de técnicas de sumarização extrativa de textos, sendo sua avaliação realizada através da verificação dos seguintes elementos: coerência, preservação da idéia central e capacidade de gerar categorias corretas. Todos esses elementos foram avaliados para diferentes taxas de compressão atribuídas aos documentos fonte.

O metadado *Subject* foi obtido através da técnica de categorização assistida, onde primeiramente o Modelo foi “ensinado” a identificar características específicas das categorias previstas (Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural). Uma vez gerados os vetores de categorias, adotou-se uma estratégia aqui denominada de *Categorização por Diferença de Similaridade*, para atribuição das categorias aos documentos. A sua avaliação se baseou na taxa de acerto da geração das categorias dos documentos. Documentos de categorias não “ensinadas” ao Modelo também foram processados para verificar seu comportamento.

Apesar de não ser objetivo principal dessa Tese a avaliação de desempenho dos algoritmos de *stemmer* Porter e RSLP utilizados nesse experimento, estes algoritmos foram

avaliados em termos de influência na geração dos metadados *Description* e *Subject*. Vale a pena ressaltar que não se encontrou trabalhos que tenham realizado avaliações de algoritmos de *stemmer* no âmbito da categorização e da sumarização de textos. A qualidade dos *stemmings* gerados também foi avaliada.

1.4 Contribuições

Primeiramente, o trabalho procura incentivar as comunidades envolvidas na produção, gerenciamento, distribuição, uso e preservação de recursos de informação, a seguirem o caminho da migração para ambientes digitais semânticos. Essa migração é um desafio de grande interesse, uma vez que a tecnologia digital, juntamente com a anotação semântica dos recursos, tem o potencial de assegurar a longevidade da informação de pesquisa; aperfeiçoar buscas e descobertas; prover acessos efetivos; oferecer compartilhamento, integração e reuso de informação; proteger direitos autorais; oferecer disponibilidade global etc.

Ao manter impulso em tais iniciativas, essas comunidades incentivarão e darão suporte a uma economia e sociedade baseadas em conhecimento, reforçando os valores culturais e sociais do país, além de nutrir oportunidades para inovação, riqueza e geração de trabalho.

O trabalho também contribui para um melhor entendimento e otimização das atividades envolvidas na construção da Web Semântica, que persegue um caminho mais realista e adequado às novas tecnologias e às novas exigências de usuários.

De forma mais detalhada, esta Tese traz as seguintes contribuições:

- a) Conjunto teórico amplo de assuntos relacionados à criação de ambientes digitais semânticos;
- b) Apresentação de um Modelo de geração automática de metadados baseado em contexto, a partir de documentos textuais na língua portuguesa, em formato não estruturado (txt). Tais metadados seguem o padrão internacional *Dublin Core*;
- c) Definição de um processo automático para conversão de documentos textuais na língua portuguesa, no formato não estruturado, em páginas *Web* anotadas semanticamente, dentro dos padrões internacionais RDF/XML e *Dublin Core* recomendados pela Web Semântica. A adoção desses padrões facilitará o compartilhamento e reuso das informações tanto por usuários como por programas de computador;

- d) Análise do comportamento dos algoritmos de *stemmer* Porter e RSLP na categorização e sumarização de documentos textuais na língua portuguesa;
- e) Definição de uma estratégia de atribuição de categorias a documentos: *Categorização por Diferença de Similaridade*;
- f) Análise do comportamento do Modelo proposto;
- g) Disponibilização do Modelo proposto para novos experimentos.

1.5 Trabalhos Relacionados

Abaixo se encontram descritos os principais trabalhos relacionados diretamente com a geração automática ou semi-automática de metadados:

- a) Prospecção de Dados na Manutenção Semi-automática de Ontologias.

Este trabalho [Barbosa2003] desenvolveu uma aplicação de manutenção semi-automática de uma ontologia de domínio de uma comunidade científica. Para tanto, foi utilizado um portal Internet que incorporou uma ontologia como alicerce semântico que conecta os fornecedores aos consumidores de conhecimentos disponíveis, aprimorando sua funcionalidade. Para realizar tal tarefa, executaram-se os seguintes passos:

- Análise da ontologia de domínio incorporada ao portal Internet;
- Identificação do conhecimento necessário para a manutenção da ontologia;
- Análise de modelos de prospecção de uso da *Web* objetivando a prospecção de dados apropriados à obtenção deste conhecimento;
- Projeto e desenvolvimento de técnicas de prospecção de uso da *Web* para descoberta e análise do conhecimento que foi utilizado por especialistas na manutenção semi-automática da ontologia;
- Incorporação das técnicas desenvolvidas ao portal Internet escolhido.

Para mostrar a viabilidade da proposta foi implementada parte da arquitetura para uma aplicação de Gestão de Conhecimento envolvendo uma comunidade científica possuidora de um portal Internet disponibilizando serviços, para o qual foi desenvolvida uma

ontologia do domínio específico de conhecimento, que representa um mapa do conhecimento da comunidade envolvida. A idéia é que a partir de uma determinada necessidade de informação, o visitante acesse o portal *Web* e siga trilhas de navegação guiadas pelas suas preferências e necessidades de informação. As trilhas percorridas através dos conteúdos/recursos do portal são automaticamente registradas nos arquivos de *logs*, sendo preparadas e agrupadas para submissão a procedimentos de prospecção de dados, que gerarão conhecimento útil para a manutenção da ontologia e dos perfis dos usuários.

A partir dos termos utilizados nas consultas realizadas pelo usuário, juntamente com os dados do contexto em que estes termos foram selecionados, o sistema gera perfis dos usuários, possibilitando a melhoria da cooperatividade e desempenho das consultas. Isso tem o objetivo de obter os benefícios provenientes da personalização, como a habilidade de servir às necessidades dos consumidores e apresentar o melhor e mais apropriado serviço para clientes individuais a qualquer momento. Por outro lado, tais dados, juntamente com os dados de resultados retornados e utilizados pelos usuários a partir das consultas guiadas pela ontologia, foram utilizados para o sistema fazer sugestões para a manutenção da ontologia.

O sistema foi baseado em agentes de *software* que apóiam as interações do usuário com o portal, filtrando as informações que lhes são relevantes e que apoiarão os engenheiros do domínio na tomada de decisão no processo de manutenção da ontologia.

b) Automatic Metadata Discovery from non-cooperative Digital Libraries site.

Este trabalho [SMZ2003] descreve a arquitetura de um serviço unificado de descoberta automática de metadados, para um conjunto heterogêneo de bibliotecas digitais que possuem suas próprias regras para prover, obter, definir e exibir metadados em páginas HTML.

O serviço permite a um usuário procurar, de forma transparente, através de um domínio de páginas *Web* “escondidas” de bibliotecas digitais, e obter um conjunto de metadados tão grande quanto possível, que são armazenados em um repositório local, criando um *cache* inteligente. Este *cache* oferecerá melhor desempenho e eficiência para as futuras pesquisas do usuário (pois nesse caso primeiramente serão checadas as informações locais, para depois enviar consulta às bibliotecas digitais remotas).

Em geral, bibliotecas digitais provêm serviços de busca com interfaces *Web* que oferecem uma página de pesquisa, uma lista de resultados da pesquisa, e página com

documento/registro individual detalhado. Tanto a lista de resultados da pesquisa como a página detalhada de documento/registro é fonte possível de metadados. Assim, este trabalho definiu regras para extrair tais metadados e desenvolveu uma análise gramatical de metadados que usam essas regras para obter os metadados. Foi utilizado o conjunto de elementos de metadados da *Dublin Core* para mapear cada campo de metadado das bibliotecas digitais.

c) MetaSIG: Ambiente de Metadados para Aplicações de Sistemas de Informações Geográficas.

O MetaSIG [PTC1999] é um ambiente de metadados que possui como maior objetivo a extração automática de esquemas e metadados presentes em aplicações geográficas, independente do modelo utilizado. Para alcançar tal objetivo, foi utilizada uma “arquitetura de quatro camadas” para meta modelagem, comumente utilizada para descrever os diferentes níveis de informação dentro do domínio de modelagem. Tais camadas foram as seguintes:

- Camada “Dados do Usuário”: refere-se a um domínio de informação específico. Dentro do escopo de SIG (Sistema de Informação Geográfica), pode-se considerar como instâncias desta camada os dados de aplicações geográficas armazenados em Bancos de Dados Geográficos de ambientes de SIG;
- Camada “Esquemas”: contém os esquemas conceituais resultantes da aplicação de um modelo, sobre um determinado domínio de informação;
- Camada “Meta-esquemas”: contém os esquemas conceituais cujos domínios de informação referem-se, especificamente, ao modelo de dados, sejam eles modelos conceituais ou físicos. Esta pode ser considerada como sendo o “Nível dos Modelos”;
- Camada “Metameta-esquema”: define a linguagem para especificar meta-esquemas; ou seja, através de um determinado metameta-esquema deve ser possível criar instâncias de meta-esquemas, como por exemplo, os meta-esquemas dos Modelos Conceituais Geográficos.

Seguindo a hierarquia da arquitetura de quatro níveis, pode-se considerar “Dados do usuário” como sendo uma instância de “Esquemas”, “Esquemas” como sendo uma instância

de “Meta-esquemas”, e, finalmente, “Meta-esquemas” como sendo uma instância de “Metameta-esquema”.

O MetaSIG é capaz de extrair, automaticamente, metadados de dois tipos de fontes diferentes:

- Metadados provenientes de “catálogos” de Bancos de Dados Geográficos utilizados por ambientes de SIG;
- Metadados provenientes de “Bancos de Dados” utilizados por ferramentas de projeto de aplicações geográficas.

Devido à sua capacidade de extrair metadados a partir de fontes heterogêneas, o MetaSIG pode ser considerado um ambiente de metadados genéricos. Para isso, ele se utiliza dos conceitos de meta-esquema e metameta-esquema já descritos anteriormente.

O responsável pela confecção do meta-esquema dentro do Ambiente MetaSIG é o que se denomina “Usuário Integrador”, também chamado de “*broker*”. Este usuário, além de possuir conhecimentos sobre modelagem de dados, deve conhecer o modelo da fonte de metadados, para, junto ao usuário de SIG, levantar os metadados que irão compor o meta-esquema. Estes metadados podem ser automaticamente extraídos da fonte ou ainda podem ser metadados complementares a serem introduzidos pelo próprio usuário de SIG.

d) Automatic RDF Metadata Generation for Resource Discovery.

Este trabalho [JJBW1998] da *University of Wolverhampton* propõe que metadados de descoberta de recursos sejam gerados a partir do processo de classificação automática de documentos HTML definido para a WWLib (*Wolverhampton Web Library*).

O trabalho apresenta um conjunto de elementos de metadados (*Wolverhampton Core*) que foram definidos como sub-propriedades das propriedades definidas para o padrão *Dublin Core*. Durante a análise gramatical dos documentos HTML a serem classificados, termos encontrados no elemento *<TITLE>*, palavras-chave e descrições encontradas no META TAG *Description* são extraídos facilmente, e atribuídos pesos. Outros elementos extraídos são: número de acesso, URL, classificação, número de palavras, data da classificação, data da última modificação. Tais metadados são representados em RDF.

O classificador automático recupera documentos HTML de determinadas URLs, analisa seus conteúdos e atribui classes DDC (*Dewey Decimal Classification*) apropriadas. A hierarquia de classes DDC foi definida manualmente.

A Tabela 1.3 exibe uma comparação dos trabalhos relacionados descritos acima, com o trabalho aqui proposto:

Tabela 1.3 – Tabela Comparativa dos Trabalhos Relacionados e Trabalho Proposto

	Objetivo	Fonte dos Metadados	Tipo de Extração	Categoria	Padrão	Produto Final
Barbosa 2003	Manutenção semi-automática de ontologia e perfil de usuário.	Arquivos de <i>log</i>	Semi-automática: interferência dos especialistas da ontologia	Utilização	Não especificado	Portais Internet com funcionalidades aprimoradas
SMZ 2003	Serviço unificado de descoberta automática de metadados	Páginas HTML e lista de resultados de pesquisas	Automática	Descritivo	<i>Dublin Core</i>	BDs de metadados para oferecer melhor desempenho em sistemas de RI
PTC1999	Extração automática de esquemas e metadados presentes em aplicações geográficas (SIG)	Catálogos de BD geográficos de ambientes de SIG; BD de ferramentas de projeto de aplicações geográficas.	Automática ou semi-automática (auxílio do usuário do SIG)	Descritivo	Não especificado	BDs de metadados para os diversos meta-esquemas
JJBW 1998	Geração de metadados de recuperação de recursos	Documentos textuais em HTML da língua inglesa	Automática	Descritivo	<i>Wolverhampton Core</i> - subconjunto do <i>Dublin Core</i> .	Metadados em RDF/XML
Trabalho Proposto	Geração de metadados de gerenciamento e recuperação de recursos	Documentos textuais não estruturados, na língua portuguesa	Automática e semi-automática	Administrativo, descritivo e técnico	<i>Dublin Core</i>	Páginas Web anotadas semanticamente em RDF/XML/DC

De um modo geral, as diferenças básicas entre os trabalhos relacionados e a proposta dessa Tese estão principalmente nos itens *Fonte dos metadados* e *Categoria* da Tabela 1.3.

A fonte dos metadados, que nos trabalhos relacionados são arquivos estruturados ou semi-estruturados (como banco de dados, arquivos de *log* e documentos HTML), apresentam alguma estrutura que explicitam a localização e o conteúdo da informação. O trabalho aqui proposto extrair metadados de documentos não estruturados na língua portuguesa.

Com relação à categoria dos metadados gerados, o trabalho aqui proposto gera metadados descritivos, administrativos e técnicos.

1.6 Organização do Trabalho

Esta Tese é composta por sete capítulos um apêndice e dois anexos, cujos objetivos estão descritos abaixo:

- Capítulo 1 – *Introdução*.
Apresenta os objetivos, justificativas, metodologia, contribuições, trabalhos relacionados e organização do trabalho.
- Capítulo 2 – *Ambientes Digitais e a Web Semântica*.
Apresenta os conceitos básicos de ambientes digitais. Define a *Web Semântica* e seus componentes principais: XML, RDF, Ontologias, metadados e agentes.
- Capítulo 3 – *Metadados: apenas Dados sobre Dados?*
Define metadados, sua importância, categorias, funções e padrões internacionais.
- Capítulo 4 – *Descoberta de Conhecimento em Texto*.
Apresenta as principais técnicas de Inteligência Artificial para descoberta de conhecimento, mais especificamente na área de Mineração de Texto, bem como sobre tecnologias de Recuperação de Informação (RI) e de Processamento de Linguagem Natural (PLN). Apresenta também os principais conceitos envolvidos nas tarefas de categorização e sumarização automática de textos.
- Capítulo 5 – *SIGAMeta - Um Modelo de Geração Automática de Metadados*.
Este capítulo define a arquitetura geral do Modelo e seu funcionamento, interfaces gráficas, o conjunto de elementos de metadados, métodos e algoritmos para extração dos metadados, e métricas e métodos de avaliação.
- Capítulo 6 – *Avaliação e Análise dos Resultados*.
Apresenta os resultados obtidos e faz avaliações e análises.
- Capítulo 7 – *Conclusões, Limitações e Trabalhos Futuros*.
Apresentam as conclusões do trabalho, suas limitações e relaciona os possíveis trabalhos futuros.
- Apêndice A – *Desafios na Construção e Manutenção de um Ambiente Digital*

Apresenta alguns dos principais desafios enfrentados no desenvolvimento de ambientes digitais voltados para instituições centradas em gestão e disseminação da informação.

- Anexo A – *Diagramas Comportamentais do Modelo SIGAMeta*

Apresenta o Diagrama de Caso de Uso e o Diagrama de Seqüência, em linguagem UML, obtidos através da ferramenta *Poseidon for UML – Community Edition*.

- Anexo B – *Informações e Resultados Complementares*

Apresenta listas e tabelas complementares de resultados obtidos nos experimentos.

Ambientes Digitais e a Web Semântica

2.1 Ambientes Digitais Voltados à Gestão e Disseminação de Informação

A crescente informatização da sociedade e a percepção da informação como valor, são alguns dos fatores que têm colaborado para consolidar um novo conceito a respeito de instituições centradas em gestão e disseminação de informação, como bibliotecas, museus, arquivos, instituições de ensino e pesquisa, comunidades de preservação, comunidades científicas e culturais. Não mais restritos a quatro paredes, nos últimos anos tais ambientes vêm passando por modificações radicais, obedecendo a um novo ritmo de funcionamento, permanentemente disponível e sem fronteiras. São as bibliotecas, museus e arquivos virtuais, também chamados digitais ou eletrônicos, disponíveis na Internet.

Em virtude de não se ter encontrado, em trabalhos anteriores, uma definição de “ambientes digitais” no contexto adotado aqui nesta Tese, tomou-se a liberdade de defini-los conforme se encontra abaixo. Tal definição foi formulada a partir de idéias e conceitos atribuídos a bibliotecas e museus digitais, adotados pela *Digital Library Federation* e pela *National Science Foundation* [Cleve1998, Borg1999].

Ambientes digitais voltados à gestão e disseminação de informação não são apenas uma coleção de informação digitalizada. Eles são instituições que envolvem serviços e pessoas, e encerram sistemas extremamente complexos de informação que dão suporte à criação, estruturação, gestão, pesquisa e recuperação, uso, distribuição e preservação de objetos de informação. Manipulam dados em qualquer formato (texto, imagens estáticas ou dinâmicas, sons etc.) e existem em redes distribuídas. O conteúdo dos ambientes digitais inclui dados e metadados que refletem conteúdo, contexto e estrutura dos objetos de informação. Os ambientes digitais são construídos para apoiar comunidades específicas de usuários, e suas capacidades funcionais dão suporte às necessidades da informação e ao uso daquela comunidade. Esses ambientes permitem que indivíduos, grupos ou programas interajam entre si, compartilhando dados, informações, recursos de conhecimento e sistemas.

Dependendo da finalidade do ambiente digital (biblioteca, museu, arquivo etc.) e da comunidade servida por esse ambiente, os serviços oferecidos podem ser mais ou menos especializados. Por exemplo, uma instituição voltada a arquivamento de informação digital provavelmente desenvolverá um projeto de preservação de seus acervos bem mais sofisticado que uma instituição de ensino e pesquisa. Bibliotecas digitais, por sua vez, adotarão técnicas sofisticadas de recuperação e visualização de recursos digitais.

Os ambientes digitais podem ser justificados pelos seus benefícios e vantagens, quando comparados com seus respectivos ambientes tradicionais [GF2001]. Por exemplo:

- 1) Facilidade de atualização das informações e alta qualidade dos recursos multimídia, ajudando a remover barreiras físicas e conceituais;
- 2) Disponibilidade global e contínua das informações, além de conectividade de rede e tecnologias interativas. Tudo isso propicia interações sociais que permitem a criação de espaços de trabalho virtuais ricos;
- 3) Uso da tecnologia digital que permite a construção de serviços avançados e inovadores, muitos dos quais com características não possíveis ou extremamente difíceis ou caros de serem implementados em ambientes tradicionais. Estes incluem, por exemplo, máquinas de procura poderosas e efetivas, novos métodos de visualização e interação com a informação, e disseminação automática da informação;
- 4) Preservação dos objetos originais de informação.

O processo de transição dos ambientes convencionais para ambientes digitais tem suscitado diversas questões em torno de assuntos como direitos autorais, extinção de funções e serviços (como o bibliotecário), extinção de mídias tradicionais físicas (como o suporte de papel) e outros. Em [Cleve1998, Borg1999] encontram-se algumas observações a respeito dessas questões no âmbito das bibliotecas digitais, que aqui foram reformuladas para caracterizar ambientes digitais voltados à gestão e disseminação de informação:

- Os ambientes digitais, além de exibirem o lado digital dos ambientes tradicionais, incluem também objetos de informação em suporte tangível, ou representados por

entidades analógicas, ou ainda objetos digitais em formato não padrão. Por exemplo: arquivos originais em formato eletrônico (textos, planilhas, gráficos, imagens e sons), objetos textuais sobre papel (livros, folhetos, manuscritos), objetos que contenham informação gráfica bidimensional (fotografias, microfilmes, negativos, selos), objetos tridimensionais (pinturas, estátuas, esculturas, edifícios), gravações sonoras em discos e fitas, gravações audiovisuais em bitolas e fitas VHS etc. Todos esses formatos não são excludentes, assim como não o são os diferentes veículos de comunicação (rádio e TV, vídeo e cinema);

- Os ambientes digitais incluem todos os processos e serviços que compõem seus respectivos ambientes tradicionais, porém, estes deverão ser revisados e aperfeiçoados para acomodarem as diferenças entre as novas mídias digitais e as mídias fixas tradicionais. O perfil do profissional de ciência da informação vem se redefinindo de acordo com a incorporação de novas tecnologias específicas desta área, acompanhando também as mudanças mais amplas ocorridas no mercado de trabalho, com as exigências próprias dos novos modelos econômicos e de gestão do trabalho;
- Os ambientes digitais idealmente provêm uma visão coerente de tudo sobre as informações contidas em seus domínios, não importando a forma ou formato das informações;
- Os ambientes digitais servem a comunidades particulares ou grupos, conforme seus ambientes tradicionais atuais, apesar de que essas comunidades podem estar dispersas amplamente ao longo da rede;
- Os ambientes digitais necessitam tanto das habilidades de funcionários dos ambientes tradicionais, como também dos cientistas de computador, para serem viáveis.

Na polêmica em torno dos direitos autorais, questiona-se, entre outras coisas, a facilidade proporcionada pelos objetos de informação digital de reproduzir e distribuir cópias não autorizadas, e a ausência de uma legislação mais específica. A lei 9.610 de 1998 sobre direitos autorais no Brasil [Lei1998] também é válida para a Internet. De acordo com ela, os direitos patrimoniais do autor, sejam para livros, obras audiovisuais ou fotográficas, duram por 70 anos após sua morte. Após este prazo, são de domínio público, podendo ser reproduzidos e copiados. Durante a vida do autor qualquer reprodução deve ser autorizada

por ele ou por terceiros, caso os direitos tenham sido concedidos para outras pessoas. Um dos fatores que dificulta a criação de uma legislação direcionada especificamente para esta questão, é a territorialidade das leis de direitos autorais e a queda de fronteiras caracterizada pela Internet.

Segundo a *Digital Library Federation* e a *National Science Foundation* [Cleve1998, Borg1999], devido à complexidade envolvida na construção de bibliotecas digitais, estas com certeza estão longe de serem sistemas únicos e completos que provêem acesso imediato a todas as informações, para todos os setores de sociedade, em qualquer lugar no mundo. Ao contrário, bibliotecas digitais certamente permanecerão sendo coleções de recursos e sistemas discrepantes, abastecendo comunidades específicas e grupos de usuário, criados para propósitos específicos.

A afirmação acima é aceitável considerando, por exemplo, a necessidade das bibliotecas, ou qualquer outro ambiente digital, de captarem as infinitas perspectivas de usuários diante de uma mesma informação. Além do mais, dificuldades relativas à interoperabilidade entre ambientes digitais (de arquiteturas técnicas, metadados e formatos de documentos) também contribuem significativamente para que se acredite que estes só poderão existir dentro de sistemas relativamente limitados desenvolvidos para propósitos e comunidades específicas.

Entretanto, novas alternativas aparecem com o conceito de *Web Semântica* e suas tecnologias adjacentes, que segundo seu idealizador Tim Berners-Lee [BLH2001], será uma extensão da *Web* atual, porém apresentará estrutura que possibilitará a compreensão e o gerenciamento dos conteúdos armazenados na *Web*, tanto por pessoas como por programas, independente da forma em que estes se apresentem (texto, som, imagem e gráficos). Isso tudo será possível a partir da valoração semântica dos conteúdos dos documentos digitais, e através de agentes que são programas coletores de conteúdos advindos de fontes diversas, capazes de processar as informações e permutar automaticamente resultados com outros programas.

Assim sendo, as tecnologias envolvidas na Web Semântica certamente impulsionarão o desenvolvimento de arquiteturas de ambientes digitais interoperáveis e mais inteligentes, viabilizando a geração de sistemas mais completos que atenderão às necessidades de comunidades cada vez mais amplas e heterogêneas.

2.2 A Web Atual e a Web Semântica

A forma mais usual de disponibilizar e acessar a informação digital tem sido através da *Web* (também conhecida como *WWW*), que buscava oferecer interfaces mais amigáveis e intuitivas para a organização e acesso ao crescente repositório de documentos que se tornava a Internet.

Atualmente, estando a *Web* com mais de 1 bilhão de páginas indexáveis (ver <http://www.inktomi.com/webmap/>), seus usuários enfrentam sérios problemas na recuperação da informação desejada. Em sua grande maioria, ferramentas automatizadas para recuperação de informação utilizam esquemas de indexação baseados em palavras-chave contidas nos textos, o que é muito pouco eficaz quando não se tem uma estrutura de dados que se explicita o contexto das informações. Além disso, os mecanismos de busca só conseguem dar uma cobertura limitada e pequena das informações contidas na *Web*, de forma que a grande maioria de seus documentos permanece encoberta [SA2004].

A dificuldade de se explicitar o contexto dos documentos disponíveis na *Web* tem exigido do usuário uma análise mais profunda dos resultados de suas buscas. Além disso, a ênfase das tecnologias e linguagens atualmente utilizadas nas páginas *Web* (como o HTML) focaliza os aspectos de apresentação dos dados, sem expor o seu significado, de forma que a informação se torna pobremente descrita e pouco passível de ser consumida por máquinas e seres humanos. Vem ainda se juntar a esses problemas o enorme crescimento desta rede e a ampliação das possibilidades de sua utilização, apontando a necessidade de uma nova filosofia com novas tecnologias, além de ampliação da infra-estrutura tecnológica de comunicação.

A *Web Semântica* vem se apresentando como solução para ordenar os problemas informacionais existente na *Web*. O objetivo deste projeto é melhorar as potencialidades da *Web* através da criação de ferramentas e de padrões que permitam atribuir significados claros aos conteúdos das páginas e facilitar a sua publicação e manutenção. As páginas construídas usando a filosofia da *Web Semântica* passam a fazer parte de um meio universal para a troca de informação, automação, integração e reuso de dados, uma vez que estas podem ser lidas por humanos ou por máquinas e tanto podem ser apresentadas graficamente (por exemplo, em monitores ou impressas em papel) como também transmitidas por sintetizadores de voz, apresentadas em equipamentos móveis (como celulares) etc.

A implementação da *Web Semântica* está baseada no RDF (*Resource Description Framework*) [W3C-RDF], que usa notação XML (*eXtensible Markup Language*) [W3C-XML] como sintaxe de codificação para descrição de metadados (combinação conhecida como RDF/XML). Outras tecnologias envolvidas são metadados, ontologias e agentes (todas apresentadas na próxima seção).

Seu funcionamento geral e a forma como suas tecnologias se relacionam pode ser melhor compreendido ao dividi-la em duas áreas de atuação: *representação e indexação de documentos*, e *recuperação e uso de documentos* [SA2004], conforme pode ser visto no *roadmap* da Web Semântica [SemanticWeb], e na Figura 2.1:

Figura 2.1 – O Roadmap da Web Semântica [SemanticWeb]

* Representação e Indexação de Documentos →

* Recuperação e Uso dos Documentos →

No âmbito da *representação e indexação dos documentos* (ver setas e números na Figura 2.1), têm-se as ferramentas e tecnologias para anotação semântica das páginas *Web* (*Web-Page Annotation Tools* - 1) e para construção de ontologias compartilhadas (*Ontology Construction Tools* - 2). Estas ferramentas possibilitarão a existência cada vez mais ampla e

disseminada de páginas *Web* marcadas semanticamente (*Annotated Web-Pages* - 3) com metadados descritos em *namespaces* de domínio público (*Metadata Repository* - 4) e com conteúdo semântico compartilhado em seu significado pelas comunidades e usuários da *Web*, através das ontologias (*Ontologies* – 5). As ontologias criadas serão articuladas entre si por meio de ferramentas específicas e meta-ontologias (*Ontologies Articulation Toolkits* - 6). Com uma estratégia padronizada de indexação, pode-se projetar sistemas mais funcionais para recuperação da informação armazenada.

No âmbito da *recuperação e uso dos documentos*, os agentes (*Agents* – 7), associados aos mecanismos de busca e inferência (*Inference Engine* - 8) executarão a coleta de informações nos documentos anotados semanticamente (3) de maneira eficaz, porque serão capazes de “compreender” seus conteúdos, de modo que a informação seja mais significativamente utilizada pelos usuários (humanos e não humanos) da *Web*. Por sua vez, os usuários poderão acessar estas novas tecnologias por meio dos portais comunitários (*Community Portals* - 9) ou mesmo pelos portais corporativos das organizações.

As infra-estruturas da Internet e das Intranets serão os dispositivos responsáveis pelo armazenamento e os canais por onde ocorrerá a disseminação dos documentos, neste grande ambiente digital semântico.

2.3 Componentes Fundamentais da Web Semântica

No âmbito da *representação e indexação de documentos*, que é o que se endereça esta Tese, a implementação da Web Semântica se baseia principalmente nas tecnologias RDF, XML, metadados e ontologias. As próximas seções apresentam estas tecnologias, com exceção do item metadados, ao qual será dedicado um capítulo exclusivo.

Figura 2.2 – Arquitetura em Camadas da Web Semântica [BLH2001]

2.3.1 O XML - *eXtensible Markup Language*

A linguagem de marcação de texto mais utilizada na construção de páginas *Web* tem sido, até hoje, o HTML (*Hiper Text Transfer Protocol*). Entretanto, como sua estrutura foi desenvolvida de uma forma muito rígida e limitada, várias dificuldades têm sido enfrentadas no atendimento de novas necessidades de usuários, como por exemplo, o entendimento do significado contextual dos documentos. Seus comandos de marcação (*tags*) foram pré-definidos para tratar unicamente da formatação dos dados a serem exibidos, e sua estrutura oferece dificuldades na adição de novos comandos de marcação que possam ampliar suas funções. A Figura 2.3 apresenta um exemplo de um documento codificado em HTML.

```
<HTML>
  <HEAD>
    <TITLE>Especificação de Microcomputador</TITLE>
  </HEAD>
  <BODY>
    <H1>Modelo</H1>
    Microcomputador Pentium IV, 1.5 GHz,
    256MB de RAM, Monitor 17 polegadas,
    Mouse, teclado, estabilizador.
  </BODY>
</HTML>
```

Figura 2.3 – Exemplo de um Documento em HTML

No exemplo acima, nota-se que a informação é um texto plano, sem indicações a que ela se refere, pois suas *tags* não explicitam o significado de seus conteúdos, e sim formato (H1). Algumas *tags* na seção <HEAD> desempenham o papel de metadados (no exemplo acima, a *tag* TITLE). Entretanto essas *tags* são poucas e mal utilizadas.

A partir das limitações do HTML e das novas necessidades dos usuários, o W3C, grupo de padronização responsável pela manutenção e avanços do HTML e outros padrões relacionados com a *Web*, definiu o padrão XML (1996). O XML é uma versão simplificada do SGML (*Standard Generalized Markup Language*), desenvolvido para ser utilizado exclusivamente na *Web*. Da mesma forma que o HTML, o XML é um padrão aberto e independente de plataforma, e não é uma linguagem de programação, e sim de marcação de texto.

O objetivo principal do XML é fornecer aos desenvolvedores da *Web* a possibilidade de criar seus próprios marcadores e atributos, à medida que estes forem necessários, bastando para isso que estejam descritos em um DTD específico (*Document Type Definition*). Em outras palavras, qualquer comunidade de desenvolvedor pode criar suas marcações específicas que atendam aos propósitos de descrição de seus dados.

Os comandos de marcação do XML estruturam o documento dividindo-o em partes identificáveis (conteúdo, significado e apresentação), de forma que pode ser indexado com maior precisão que as páginas planas escritas em HTML. Conseqüentemente, oferece uma maior flexibilidade de pesquisa. *Scripts* ou programas podem fazer uso desses comandos de marcação de forma sofisticada. Todas essas versatilidades possibilitam embutir semântica em documentos *Web*, tornando o significado de seu conteúdo melhor compreendido tanto para seres humanos como para computadores. A Figura 2.4 apresenta as mesmas informações do exemplo anterior (em HTML) codificado em XML.

```
<?xml version="1.0"?>
<MICROCOMPUTADOR>
  <MODELO>Pentium IV</MODELO>
  <VELOCIDADE>1,5 GHz</VELOCIDADE>
  <RAM>256MB</RAM>
  <MONITOR>17 Polegadas</MONITOR>
  <TECLADO>Sim</TECLADO>
  <MOUSE>Sim</MOUSE>
  <ESTABILIZADOR>Sim</ESTABILIZADOR>
  <IMPRESSORA>Nao</IMPRESSORA>
</MICROCOMPUTADOR>
```

Figura 2.4 - Exemplo de um Documento em XML

Apesar de um documento XML ser bem parecido com um HTML (pois possui *tags* para marcar as informações), o XML cria uma estrutura que demonstra claramente o significado das informações. No exemplo acima, pode-se notar que não apenas o leitor, mas também um *software*, pode interpretar corretamente as informações através das *tags* MICROCOMPUTADOR, MODELO, VELOCIDADE etc.

A estrutura lógica interna de documentos XML assume uma forma hierárquica onde cada nó dessa hierarquia é endereçável. A Figura 2.5 exibe uma parte da estrutura do documento XML acima, onde novas *tags* foram criadas para o elemento MONITOR.

Figura 2.5 – Estrutura Lógica Interna de Documentos XML

Uma vez que o XML permite a criação de *tags*, é necessário que se especifiquem regras para elaboração de documentos. Essas regras são definidas atualmente através do XML *Schema* ou do DTD (*Document Type Definition*).

O XML *Schema* e o DTD oferecem meios para verificar se um determinado documento XML contém somente os elementos e atributos que estão de acordo com a família de documentos a que pertence. Eles definem as *tags* que são válidas e suas estruturas hierárquicas, além de tipos, formatos e regras de validade de conteúdos de cada *tag*. Abaixo se encontra um exemplo de um DTD para uma agenda.

```
<!ELEMENT AGENDA (DIA+)>
<!ELEMENT DIA (HORA+)>
<!ELEMENT HORA (#PCDATA)>
<!ATTLIST AGENDA mes CDATA #REQUIRED>
<!ATTLIST DIA numero ID #REQUIRED tipo CDATA "util"
semana (segunda|terca|quarta|quinta|sexta|sabado|domingo)
#REQUIRED>
<!ATTLIST HORA valor NMTOKEN #REQUIRED>
```

Figura 2.6 – Exemplo de um DTD para uma Agenda

No exemplo acima, um documento XML para a agenda deve:

- Possuir AGENDA como elemento *root* e pode possuir vários elementos com o nome DIA;
- O elemento DIA pode conter vários elementos com o nome HORA;
- O elemento HORA pode possuir conteúdo do tipo texto;
- O documento deve conter um elemento chamado AGENDA que possui um atributo obrigatório chamado mes, que pode assumir como conteúdo um texto qualquer;
- O elemento DIA possui 3 atributos:
 - numero, que pode assumir um valor único para cada elemento;
 - tipo, que pode assumir um conteúdo texto qualquer cujo *default* é “util”;
 - semana, que pode assumir um dos conteúdos entre parênteses e também é obrigatório;
- O elemento HORA possui um atributo com o nome valor que pode assumir como conteúdo um tipo NMTOKEN válido (letras, algarismos, -, _, ., :).

Dessa forma, vê-se que o XML e o XML *Schema* (ou ainda o DTD) definem a sintaxe de um documento.

Os documentos XML não contêm em si as diretivas para exibição dos dados. Para tanto foi definida uma outra linguagem conhecida como XSL (*eXtensible Stylesheet Language*) ou “linguagem de estilo extensível” que apresenta o conteúdo de documentos XML (e também HTML). Outra opção é a CSS (*Cascading Style Sheets*), ou “folha de estilo em cascata”.

A transformação de um documento XML em um documento XSL resulta em um texto passível de ser exibido ao usuário ou entendido e utilizado por outro dispositivo tecnológico, como celulares e PDAs. No Anexo B encontra-se um documento XSL escrito para apresentação dos dados produzidos pelo Modelo proposto nesta Tese. Uma página *Web* resultante também se encontra neste Anexo.

Com a linguagem XSL, os dados contidos em um mesmo documento XML podem ser apresentados de diversas formas, onde seus elementos podem ser filtrados, exibidos segundo certas condições e ordenados segundo regras pré-definidas. Desta forma, o trio XML, seu XML *Schema* específico (ou DTD) e o XSL se apresentam como um conjunto de padrões que possibilitam o armazenamento, descrição significativa, intercâmbio e exibição dos dados de forma personalizada.

O XML é aceito como o padrão emergente para intercâmbio de dados na *Web*. Entretanto, apesar de possibilitar aos desenvolvedores a criação de seus próprios comandos de marcação (sintaxe), é necessário que estes sejam criados de forma não ambígua, para que possam ser compartilhados por comunidades maiores. Por exemplo, os comandos de marcação <CLIENTE> e <USUÁRIO> podem representar ou não a mesma coisa para desenvolvedores diferentes. Para que estas marcações sejam utilizadas de forma não ambígua, são necessários alguns padrões de compartilhamento universais. O W3C e as comunidades de usuários têm procurado prover estes padrões, como por exemplo, o padrão de metadados *Dublin Core*, abordado no Capítulo 3.

2.3.2 O RDF - *Resource Description Framework*

Como pôde ser visto na Seção 2.3.1, o XML e o XML *Schema* permitem ao desenvolvedor adicionar estruturas a seus documentos (sintaxe), conforme sua necessidade. Entretanto, o XML não explicita o que tais estruturas significam (semântica). É através do RDF que estes significados são expressos.

O RDF é uma recomendação do W3C, e constitui-se em uma arquitetura genérica de metadados compreensível por máquina, que permite descrever recursos no contexto *Web* [MM2004]. O W3C introduziu o RDF na tentativa de resolver um dos principais desafios encontrados pelas diferentes comunidades de descrição de recursos: prover interoperabilidade entre os diversos padrões de metadados, de forma a obter automação, compartilhamento, integração e reuso de informação em vários níveis e a várias aplicações.

A tecnologia RDF apóia várias áreas de aplicação, como por exemplo: descoberta de recursos, onde o RDF possibilita a implementação de mecanismos de busca mais eficientes; catalogação e classificação, onde o RDF pode ser utilizado para descrever e identificar o contexto dos recursos de informação; agentes inteligentes, onde o RDF pode facilitar a descrição e o compartilhamento do conhecimento. Outras áreas beneficiadas são: avaliação de conteúdos, direitos de propriedade intelectual, preferências de privacidade de usuários e políticas de privacidade de um *Web site*.

Tecnicamente, o RDF não é uma linguagem, mas sim um modelo de dados para descrição de recursos com maior semântica, através da adoção de metadados.

O modelo de dados RDF é simples e utiliza o conceito de “declaração” (*statement*), definido na forma de tripla composta de *predicate* (propriedade), *subject* (recurso) e *object* (valor de uma propriedade), cuja notação é (predicate,[subject],[object]). Traduzindo a notação anterior, uma declaração representa a relação entre um recurso, uma de suas propriedades e o valor que essa propriedade pode assumir.

Além do formato de tripla, o modelo RDF também pode ser visualizado na forma de um grafo, que consiste de um conjunto de nós conectados por arcos rotulados, onde os nós representam os recursos *Web* e os arcos representam as propriedades desses recursos. Os valores são representados por retângulos.

As instâncias do modelo RDF podem ser escritas usando os comandos de marcação do XML (*tags*), permitindo que suas declarações sejam analisadas sintaticamente por

analísadores XML, e passadas para aplicativos que entendam o XML. Este é um dos principais aspectos que tem contribuído para o sucesso da tecnologia RDF.

Como exemplo ilustrativo, abaixo se encontra um documento RDF/XML com suas representações em formato de tripla e de grafo. Para obtenção dessas representações, utilizou-se o *site* da W3C “<http://www.w3.org/RDF/Validator/>”. O documento RDF/XML foi obtido automaticamente pelo Modelo SIGAMeta proposto nesta Tese.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xmlns:dcterms='http://purl.org/dc/terms/'>
<rdf:Description>
  <dc:subject>
    <rdf:Bag>
      <rdf:li>Antropologia</rdf:li>
      <rdf:li>cavalo marinho</rdf:li>
      <rdf:li>folclore</rdf:li>
      <rdf:li>pernambuco</rdf:li>
    </rdf:Bag>
  </dc:subject>
  <dc:description>
    Cavalo Marinho é um auto popular que, como o bumba-meu-boi, é do
    ciclo natalino se estendendo até a Festa de Reis. Faz parte do
    folclore da região da Mata Norte de Pernambuco. É composto de
    diálogos, danças e brincadeiras, havendo, inclusive, a
    participação do público.
  </dc:description>
  <dc:publisher>FJN</dc:publisher>
  <dc:date>2006-01-19</dc:date>
  <dc:type>text</dc:type>
  <dc:format>text/xml</dc:format>
  <dc:identifier>Antrop4</dc:identifier>
  <dc:language>pt-br</dc:language>
  <dc:rights>FJN</dc:rights>
  <dc:title>Cavalo Marinho</dc:title>
</rdf:Description>
</rdf:RDF>
```

Figura 2.7 – Declaração RDF em Formato XML

Subject	Predicate	Object
genid:ARP641	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/1999/02/22-rdf-syntax-ns#Bag
genid:ARP641	http://www.w3.org/1999/02/22-rdf-syntax-ns# 1	"Antropologia"
genid:ARP641	http://www.w3.org/1999/02/22-rdf-syntax-ns# 2	"cavalo marinho"
genid:ARP641	http://www.w3.org/1999/02/22-rdf-syntax-ns# 3	"folclore"
genid:ARP641	http://www.w3.org/1999/02/22-rdf-syntax-ns# 4	"pernambuco"
genid:ARP640	http://purl.org/dc/elements/1.1/subject	genid:ARP641
genid:ARP640	http://purl.org/dc/elements/1.1/description	"Cavalo Marinho é um auto popular que, como o bumba-meu-boi, é do ciclo natalino se estendendo até a Festa de Reis. Faz parte do folclore da região da Mata Norte de Pernambuco. É composto de diálogos, danças e brincadeiras, havendo, inclusive, a participação do público."
genid:ARP640	http://purl.org/dc/elements/1.1/publisher	"FJN"
genid:ARP640	http://purl.org/dc/elements/1.1/date	"2006-01-19"
genid:ARP640	http://purl.org/dc/elements/1.1/type	"text"
genid:ARP640	http://purl.org/dc/elements/1.1/format	"text/xml"
genid:ARP640	http://purl.org/dc/elements/1.1/identifier	"Antrop4"
genid:ARP640	http://purl.org/dc/elements/1.1/language	"pt-br"
genid:ARP640	http://purl.org/dc/elements/1.1/rights	"FJN"
genid:ARP640	http://purl.org/dc/elements/1.1/title	"Cavalo Marinho"

Figura 2.8 – Representação em Formato de Tripla para um Documento RDF/XML

Figura 2.9 – Representação em Formato de Grafo para um Documento RDF/XML

Todos os objetos descritos pelo RDF são chamados de *recurso*. Estes podem ser páginas *Web*, elemento HTML ou XML específico, coleção completa de páginas etc. Objeto não acessível diretamente pela *Web*, como livros impressos, também podem ser recursos.

Os recursos são nomeados por URIs (*Uniform Resource Identifier*) ou por identificadores opcionais de âncora (*anchor*). A URI é uma forma globalizada de se nomear objetos de informação (generalização do conceito de URL).

Conforme Manola e Miller [MM2004], a especificação do RDF está dividida em duas partes principais. A primeira parte (RDF) descreve os recursos em termos de suas propriedades e valores (parte definida acima); a segunda parte, o RDF *Schema* (RDFS), é uma linguagem utilizada para descrever ontologias, pois define formalmente as relações entre termos e conceitos. Essas duas definições combinadas costumam ser referenciadas como RDF(S).

O RDF *Schema* é um mecanismo que provê a construção de estruturas hierárquicas. Ele define um sistema de classes extensível e genérico que pode ser utilizado como base

para esquemas de um domínio específico, permitindo que comunidades de descrição de recursos possam criar e compartilhar seus próprios vocabulários. Os esquemas do RDFS podem ser compartilhados e estendidos através de refinamento de subclasses. Além disso, definições de metadados podem ser reutilizadas através do compartilhamento de esquemas.

Os elementos constituintes do RDFS incluem:

- `rdfs:Resource`: todas as coisas descritas são recursos (objeto);
- `rdfs:Property`: propriedades;
- `rdfs:Class`: classe;
- `rdfs:type`: indica que um recurso é membro de uma classe. O valor será sempre um recurso que é uma instância de `rdfs:Class`;
- `rdfs:subClassOf`: declara a relação de subclasse/superclasse entre duas classes. Uma classe pode ser subclasse de mais de uma classe;
- `rdfs:subPropertyOf`: é uma instância de `rdf:Property` que é usada para especificar que uma propriedade é uma especialização de uma outra.

As figuras 2.10 e 2.11 abaixo apresentam um exemplo de grafo e o código que ilustram um RDFS.

Figura 2.10 – Grafo de uma Hierarquia de Classes RDFS

```
<rdf:RDF xml:lang="en" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

  <rdf:Description ID="VeiculoMotor">
    <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
  </rdf:Description>

  <rdf:Description ID="Carro">
    <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="#VeiculoMotor"/>
  </rdf:Description>

  <rdf:Description ID="Caminhao">
    <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="#VeiculoMotor"/>
  </rdf:Description>

  <rdf:Description ID="Van">
    <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="#VeiculoMotor"/>
  </rdf:Description>

  <rdf:Description ID="MiniVan">
    <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="#Van"/>
    <rdfs:subClassOf rdf:resource="#Carro"/>
  </rdf:Description>

</rdf:RDF>
```

Figura 2.11 – Código RDFS

Como pode ser visto na Figura 2.11, o código do RDFS é semelhante à sintaxe do RDF.

O mecanismo RDF *Schema* tem sido associado à modelagem ontológica de domínios, à medida que permite, através de um vocabulário distinto, a definição de modelos de objetos com semântica completamente definida para um domínio particular de interesse. Entretanto,

este mecanismo provê somente uma semântica estrutural, permitindo a definição de um conceito em termos de suas propriedades, das restrições impostas a estas propriedades, dos relacionamentos entre estas propriedades e dos relacionamentos com outros conceitos. A modelagem de axiomas ontológicos, responsável por prover uma maior semântica conceitual, não é contemplada pela tecnologia RDF *Schema* [SEMD2000].

2.3.3 Ontologia

Apesar do tópico Ontologia ser atualmente popular, ele parece ainda trazer um pouco de confusão sobre o que é e o seu papel na Web Semântica.

Ontologia tem sido constantemente interpretada como esquemas de classificação, taxonomias, hierarquias, *thesaurus*, vocabulários controlados, terminologias e até mesmo dicionários. Apesar das ontologias exibirem características remanescentes destes sistemas, compará-las com qualquer um tipo de estrutura de representação é diminuir sua função e seu potencial na evolução da Web Semântica [Jacob2003].

Em filosofia, ontologia é definida como sendo o “estudo dos tipos das coisas que existem” [CJB1999]. Na Ciência da Computação, esse termo vem sendo aplicado desde a década de 90 no campo da Inteligência Artificial (IA), para representação computacional de conhecimento em áreas como Engenharia de Conhecimento e Processamento de Linguagem Natural. Neste campo, o termo ontologia é aplicado a um dos dois significados descritos a seguir:

Em um primeiro sentido, ontologia é um vocabulário de representação, normalmente especializado para algum domínio ou assunto de estudo. Mais especificamente não é o vocabulário em si que se qualifica como uma ontologia, mas as conceitualizações que os termos no vocabulário pretendem capturar. Assim, ao traduzir os termos de uma ontologia de um idioma para outro, por exemplo, do Português para o Inglês, as conceitualizações da ontologia não se modificarão.

Em um segundo sentido, o termo ontologia é utilizado para referenciar um corpo de conhecimento que descreve algum domínio, tipicamente um domínio de conhecimento de senso comum, utilizando um vocabulário de representação. Em outras palavras, o vocabulário de representação provê um conjunto de termos com o qual descreve os fatos de

algum domínio, enquanto que o corpo de conhecimento que está utilizando aquele vocabulário é uma coleção de fatos sobre um domínio.

Como visto anteriormente, a estrutura atual da *Web* não oferece facilidades para o processamento automatizado de seus recursos. Como alternativa à linguagem de marcação de texto HTML, o W3C oferece o XML que vem ganhando aceitação crescente na *Web*, além do RDF que oferece um modelo de dados para descrição de recursos com maior semântica. Entretanto, esquemas XML lidam principalmente com a estrutura física de documentos (sintaxe), enquanto que o RDF(S) explicita o que tais estruturas significam. Estes, por si só, não bastam para oferecer o modelamento da semântica explícita que daria suporte à interpretação do conteúdo, de forma automática.

Se a Web Semântica tem o objetivo de proporcionar uma *Web* compreensível por máquinas, é preciso que tais máquinas saibam interpretar a semântica de cada recurso, não apenas no que diz respeito à representação precisa do conteúdo desses recursos, mas também na composição de inferências e até mesmo na descoberta de novos conhecimentos.

Dessa forma, no contexto da Web Semântica, uma ontologia é uma conceitualização parcial de um determinado domínio de conhecimento compartilhado por uma comunidade de usuários, definida em linguagem formal e processável por máquina, para o propósito explícito de compartilhar informação semântica através de sistemas automatizados [Jacob2003].

Uma ontologia oferece meios precisos e sistemáticos para definir a semântica de recursos *Web*. A ontologia especifica conceitos de domínio importantes, propriedades desses conceitos (incluindo, onde apropriado, limites de valores ou conjuntos explícitos de valores) e possíveis relações entre conceitos e propriedades.

Uma vez que uma ontologia define conceitos (os tipos de coisas e suas propriedades) e as relações semânticas que se obtêm entre esses conceitos, ela provê suporte para processamento de recursos baseado na interpretação do significado do conteúdo, ao invés de se basear na estrutura física de um recurso ou nas suas características sintáticas (como a ordenação seqüencial de ninhos de elementos).

Segundo Jacob [Jacob2003],

Qualquer esquema de metadados é, com efeito, uma ontologia que especifica o conjunto de características físicas e/ou conceituais de recursos que tenham sido julgadas relevantes

para uma comunidade particular de usuários. Dessa forma, por exemplo, o conjunto de elementos e refinamentos de elementos definidos pelo Dublin Core (DC) é por si mesmo uma ontologia.

O DC oferece consistência na descrição e interpretação da semântica de recursos da *Web*, utilizando os padrões RDF/RDFS. Apesar de oferecer uma estrutura representacional relativamente simples e aplicável para uma gama extensiva de recursos de domínio não específico, o DC é uma ontologia porque impõe um modelo conceitual formalmente definido que facilita o processo automatizado necessário para dar suporte ao compartilhamento de conhecimento entre sistemas.

Enquanto uma ontologia tipicamente define um vocabulário de conceitos de domínio dentro de uma hierarquia que apóia herança de definição de características, propriedades e limitações, o DC demonstra que esta estrutura hierárquica não é uma característica de definição de ontologias. Seus elementos são independentes um do outro: nenhum dos elementos é obrigatório pelo modelo conceitual e qualquer um pode ser repetido quantas vezes forem necessárias para qualquer recurso. Embora hierarquia não seja uma característica definida para ontologias, ela é um componente importante no modelo representacional definido pelo RDF e RDFS (como ficou mostrado na seção anterior).

Diversos padrões e linguagens para construção e compartilhamento de ontologias na *Web* estão sendo criados, todos baseados no padrão XML. Alguns exemplos são o SHOE – *Simple HTML Ontology Extensions* [SHOE], o XOL – *Ontology Exchange Language* [XOL] e o RDFS, visto na seção anterior.

Existe uma proposta de extensão do RDF e do RDFS chamada OIL – *Ontology Interchange Language* [OIL], e seu sucessor DAML+OIL – *DARPA Agent Markup Language – Ontology Interchange Language* [DAML]. Este último é uma linguagem baseada no XML, projetada para possuir muito mais capacidade que o XML na descrição de objetos e no seu relacionamento, para expressar semântica e criar um alto grau de interoperabilidade entre *sites Web*.

Sistemas de recuperação da informação, integração de fontes de informações heterogêneas e mecanismos de buscas na Internet necessitam de ontologias de domínios para organizarem informações e direcionar os processos de buscas. Por exemplo, máquinas de busca utilizam categorias e subcategorias que ajudam a organizar as buscas. Comumente, os

desenvolvedores dessas máquinas se referem a essas categorias e subcategorias como sendo ontologias. Máquinas de procura também usam ontologias para encontrar páginas com palavras que são sintaticamente diferentes, mas semanticamente semelhantes.

2.3.4 Metadados

Como visto nas seções anteriores, a dificuldade de se obter resultados precisos em buscas e descobertas na *Web* e a ênfase nos aspectos de apresentação dos dados, têm mostrado a necessidade da evolução da *Web* para uma Web Semântica. Com as ferramentas e padrões adotados pela Web Semântica, é possível atribuir significados claros aos conteúdos das páginas, e estas passam a fazer parte de um meio universal para a troca de informação, automação, integração e reuso de dados, uma vez que podem ser lidas por humanos ou por máquinas.

A automação das tarefas nesta nova *Web* se torna possível através do uso de metadados, que em sua definição mais simples significa “dado sobre dado”. Os metadados dos recursos da *Web* devem ser definidos e ligados de tal forma que seus significados possam ser interpretados não só por seres humanos, como também por programas. As tecnologias RDF e RDF *Schema* dão suporte à representação de metadados. Para que haja um entendimento comum em relação aos significados é utilizada uma ontologia.

Devido à importância dos metadados dentro dos objetivos desta Tese, o próximo capítulo detalhará este assunto.

2.4 Desafios na Construção e Manutenção de um Ambiente Digital Semântico

À medida que surgem os ambientes digitais, suas oportunidades tornam-se mais evidentes. Entretanto, também são vários os desafios a serem enfrentados em sua criação, gerenciamento e manutenção, de forma que as organizações necessitarão se posicionar estrategicamente para tirar proveito das oportunidades e tratar os desafios de forma adequada.

Alguns dos principais desafios enfrentados no desenvolvimento de ambientes digitais voltados para instituições centradas em gestão e disseminação da informação encontram-se

descritos no Apêndice A. Estes desafios foram agrupados nas seguintes categorias: provisão de acessos efetivos, construção e manutenção das coleções digitais, interoperabilidade de ambientes digitais, direitos autorais, recursos humanos e recursos financeiros para geração e sustentação das coleções.

Como pode ser visto nesse Apêndice, o trabalho aqui apresentado contribui para entender e minimizar esses desafios, especialmente no que se refere às três primeiras categorias: provisão de acessos efetivos, construção e manutenção das coleções digitais e interoperabilidade de ambientes digitais.

Metadados: apenas Dados sobre Dados?

3.1 Definição

O dado é o mais importante recurso das organizações que lidam com informação. Entretanto, para que este tenha valor, é necessário que seja entregue pontualmente, concisamente e corretamente, bem como que esteja formatado de modo a atender às necessidades dos usuários.

Os metadados são importantíssimos em ambientes digitais, visto que são os elementos chave na busca e descoberta da informação, como também no gerenciamento de todo o ciclo de vida da informação digital [Hodge2000]. Em ambientes bem planejados, valoriza-se não apenas o conteúdo das informações, mas também: as características e qualidades desses conteúdos; regras de criação, manutenção e eliminação das informações; preservação; e os meios para acessá-las.

Em sua forma mais simplista, metadados traduzem *dados sobre dados*, o que na verdade não exprime o escopo completo do que realmente representa (o termo grego *meta* significa: *depois de, após, acima de*). Segundo Gilliland [Gilliland2000], metadados é definido como sendo:

*A soma de tudo que pode ser dito sobre qualquer objeto de informação,
em qualquer nível de agregação.*

Neste contexto, um *objeto de informação* é um item, ou agregação de vários itens, de qualquer tipo ou formato (digital, analógico, suporte tangível etc.) que pode ser endereçado ou manipulado por um humano ou um sistema, como uma entidade discreta.

Tim Berners-Lee [Berners1998] apresenta uma definição semelhante, ao definir os metadados como sendo “máquina de informação compreensível sobre recursos da *Web* ou outras coisas”. Nesta definição, a expressão “máquina de informação compreensível” é fundamental, pois se refere à informação que os agentes de *software* podem usar para facilitar o trabalho dos usuários, assegurando-lhes resultado confiável e rápido.

Em geral, os objetos de informação possuem três características que podem ser refletidas através de metadados: *conteúdo*, *contexto* e *estrutura*.

- O *conteúdo* se refere a o que o objeto contém ou sobre o que trata, e é inerente ao próprio objeto de informação;
- O *contexto* indica os aspectos *quem*, *o que*, *por que*, *onde* e *como*, associados à criação do objeto, e não é inerente a um objeto de informação;
- A *estrutura* se refere ao conjunto formal de associações dentro ou entre objetos de informação individuais, e pode ser inerente ou não a um objeto de informação.

Assim, a utilização de metadados vai além de descrever e catalogar objetos de informação. Enquanto profissionais de museus, arquivos, e de biblioteca podem estar muito familiarizados com o termo em associação com descrição ou catalogação, metadados também podem indicar o contexto de criação, gerenciamento, processamento, preservação e uso dos recursos que estão sendo descritos.

É importante observar que a palavra “contexto” aqui empregada para caracterizar um metadado não corresponde ao sentido empregado frequentemente nessa Tese, quando se referindo a “contexto de documentos”. Na verdade, o que esta Tese se refere como “contexto de documentos” corresponde à característica “conteúdo” definida para metadados.

3.2 Usos e Funções

De um modo geral, os metadados têm a função de prover meios para: descoberta de recursos, disponibilidade de recursos, utilização de recursos, e administração e controle de recursos [Greenberg2003]. De forma detalhada, abaixo se encontram algumas das diversas

funções desempenhadas pelos metadados, as quais podem ser operacionalizadas por pessoas ou por agentes automatizados [Hodgson1998]:

- Fornecer um resumo do conteúdo do recurso de informação (isto é, do que trata o recurso), permitindo aos usuários determinar se este é exatamente o que eles desejam;
- Dar condições aos usuários de procurar, recuperar e usar recursos de informação;
- Quando formatos múltiplos são providos, ajudar aos usuários na escolha do recurso de informação apropriado ao seu uso;
- Prevenir usuários quanto às limitações no acesso aos dados, no que diz respeito a *softwares* necessários, padrões e formatos adotados;
- Instruir em como interpretar os dados (formato, codificação, criptografia);
- Fornecer informações que afetam o uso do recurso (por exemplo, condições legais de uso, idade mínima do usuário, cobrança de taxas, usos permitidos e não permitidos para o recurso etc.);
- Fornecer a história ou proveniência do recurso de informação, tal como sua fonte original e suas transformações subseqüentes (filtragem, atualização, descarte etc.);
- Fornecer especificações para o gerenciamento dos recursos de informação (data de última modificação, data de criação, identidade do administrador etc.);
- Fornecer especificações sobre relações entre recursos (por exemplo, o relacionamento entre um conjunto de artigos e um jornal, entre um trabalho original e sua versão subseqüente, entre componentes de um trabalho multimídia etc.);
- Listar os componentes lógicos de recursos de informação complexos e oferecer meios para acessar esses componentes (por exemplo, tabelas, lista de componentes de um *software*, etc.).

Diversas teorias e práticas são aplicadas a metadados, variando de acordo com as comunidades de profissionais que projetam, descrevem, administram, preservam e usam sistemas de informação, como por exemplo, profissionais de tecnologia da informação, de bibliotecas, de museus, de ambientes de negócios, comunidades de preservação, comunidades científicas e culturais. Essas diversas perspectivas de uso dos metadados fazem com que estes tomem formas e utilidades variadas que corresponderão às áreas específicas a

que estiverem sendo aplicados. Por exemplo, na área da biblioteconomia, os metadados são freqüentemente utilizados para descrever as fontes de seus objetos de informação, segundo determinado esquema normalizado, que permite uma recuperação uniforme e eficaz. No ambiente da Internet, os metadados foram desenvolvidos para racionalizar, diferenciar e tratar a grande quantidade de informações e, conseqüentemente, para melhorar a recuperação na *Web*.

À medida que vai amadurecendo a infra-estrutura da informação em sistemas computacionais, novos conjuntos de metadados e novas funções vão sendo desenvolvidos. O aparecimento de novas comunidades e interesses também impulsiona o surgimento de novos tipos e aplicações de metadados.

3.3 Categorias dos Metadados

Dentro de uma organização, todas as possíveis perspectivas de usuários aplicadas a metadados são importantes no desenvolvimento de uma solução de metadados. Porém, tais perspectivas ampliam enormemente o escopo de uma solução, exigindo uma maior compreensão do conceito de metadados e sugerindo quebrá-los em categorias que englobem todas as perspectivas dos usuários daquela organização. Assim, dependendo dos papéis e perspectivas dos usuários, metadados poderão ser categorizados de diferentes formas.

Gilliland [Gilliland2000] apresenta uma categorização de metadados que refletem perspectivas de gerenciamento e recuperação de recursos em ambiente digital, que é o âmbito aplicado a esta Tese. As seguintes categorias são apresentadas: metadados administrativos, descritivos, de preservação, de uso e técnicos. Conforme apresentado no Capítulo 1 dessa Tese, o Modelo aqui proposto gera metadados das categorias administrativos, descritivos, e técnicos.

A Tabela 3.1 define cada uma destas categorias e dá exemplos de funções comuns que cada uma poderia executar em um sistema de informação. Além de diferentes tipos e funções, metadados também apresentam características distintas. A Tabela 3.2 classifica os metadados segundo alguns atributos chave, e dá exemplos.

Tabela 3.1 – Categorias de Metadados que Refletem Perspectivas de Gerenciamento e Recuperação de Recursos em Ambiente Digital, e suas Funções

Tipo	Função	Exemplos
Administrativo	Metadados usados em gerenciamento e administração de recursos de informação	Informação de aquisição Registros de direitos e reprodução Critério de seleção para digitalização Controle de versão
Descritivo	Metadados usados para descrever ou identificar recursos de informação	Registros de catalogação Ajudas de descoberta Relações de <i>Hyperlinks</i> entre recursos
Preservação	Metadados relacionados ao gerenciamento de preservação dos recursos de informação	Documentação das condições físicas dos recursos Documentação das ações tomadas para preservar as versões físicas e digitais dos recursos
Técnico	Metadados relacionados a como um sistema funciona ou como metadados se comportam	Documentação de <i>hardware</i> e <i>software</i> Informação de digitalização, por exemplo, formatos, taxas de compressão Registros de tempos de resposta de sistemas Dados de segurança e autenticação, por exemplo, chaves de criptografia, senhas
Utilização	Metadados relacionados ao nível e tipo de utilização dos recursos de informação	Registros de apresentação Registros de uso e usuário Informação de reutilização de conteúdo e de multiversões

Tabela 3.2 – Atributos e Características de Metadados

Atributo	Características	Exemplos
Fonte dos Metadados	Metadados internos: gerados na ocasião em que uma informação é criada ou digitalizada pelo próprio agente criador da informação	<ul style="list-style-type: none"> - Nomes de arquivo e informação de cabeçalho - Estruturas de diretório - Formato de arquivo e esquema de compressão
	Metadados externos: gerado após a criação de uma informação, freqüentemente por alguém diferente de seu criador original	<ul style="list-style-type: none"> - Registros de inscrição e catalogação - Direitos e outras informações legais
Método de Criação de Metadados	Metadados automáticos: gerados por um computador	<ul style="list-style-type: none"> - Índices de palavras-chave - Registros de transação de usuário
	Metadados manuais: criados por pessoas	<ul style="list-style-type: none"> - Representantes descritivos tais como

		registros de catálogo e metadados <i>Dublin Core</i>
Natureza dos Metadados	<p>Metadados leigos: criados por pessoas que não são especialistas em informação nem no assunto, freqüentemente o criador original da informação</p> <p>Metadados especialistas: criados por especialistas de informação ou no assunto, freqüentemente não sendo o criador original da informação</p>	<ul style="list-style-type: none"> - <i>Metatags</i> criados por pessoal de página <i>Web</i> - Sistemas de arquivamento pessoais - Cabeçalhos de assuntos especializados - Registros MARC - Ajudas de descoberta de arquivo
Estado dos Metadados	<p>Metadados estáticos: que nunca se modificam, uma vez que eles tenham sido criados</p> <p>Metadados dinâmicos: que podem mudar com uso ou manipulação da informação</p> <p>Metadados de longo prazo: necessários para assegurar que a informação continua sendo acessível e utilizável</p> <p>Metadados de curto prazo: principalmente de natureza transacional</p>	<ul style="list-style-type: none"> - Título, proveniência e data de criação de uma informação - Estrutura de diretório - Registros de transação de usuário - Resolução de imagem - Formato técnico e informação de processamento - Informações de direitos de propriedade intelectual - Documentação de gerenciamento de preservação
Estrutura dos Metadados	<p>Metadados estruturados: que estão de acordo com uma estrutura previsível padronizada ou não</p> <p>Metadados não estruturados: que não estão de acordo com uma estrutura previsível</p>	<ul style="list-style-type: none"> - MARC - Formatos de banco de dados locais - Anotações e campos de notas não estruturados
Semânticas dos Metadados	<p>Metadados controlados: que estão de acordo com um vocabulário padronizado ou forma de autoridade</p> <p>Metadados não controlados: que não estão de acordo com qualquer vocabulário padronizado ou forma de autoridade</p>	<ul style="list-style-type: none"> - AACR2 - Notas de texto livre - <i>Metatags</i> de HTML

Nível dos Metadados	Metadados de coleção: relativos a coleções de informação	<ul style="list-style-type: none"> - Registro de nível de coleção, por exemplo, registro MARC ou ajuda de descoberta - Índice especializado
	Metadados de item: relativos a informações individuais, freqüentemente contidos dentro de coleções	<ul style="list-style-type: none"> - Legendas de imagens transcritas e datas - Informação de formato

3.4 Criação e Gerenciamento de Metadados em Ambiente Digital

Como pôde ser observado, o conceito de *metadados* envolve funcionalidades, características e atributos diversos, tornando complexos seus processos de criação e gerenciamento. Outro ponto que vem acentuar essa complexidade é o fato de diversos indivíduos e processos poderem criar e modificar metadados (de forma automática ou manual) em momentos diferentes na vida de um objeto de informação. Um modelo de ciclo de vida de objetos em sistemas de informação digital foi proposto em [Gilliland2000], onde aparecem as diversas fases que objetos de informação tipicamente se movem durante suas vidas (Figura 3.1). À medida que se movem, adquirem camadas de metadados que vão sendo associados a eles.

Projetistas de sistemas têm tido dificuldades em incorporar e gerenciar metadados em sistemas de informação, bem como em solucionar problemas relativos à manutenção destes durante o tempo de vida dos objetos associados, principalmente porque tais atividades vêm sendo predominantemente desenvolvidas manualmente.

Figura 3.1 – O Ciclo de Vida de Objetos Contidos em um Sistema de Informação Digital [Gilliland2000]

- *Criação e geração de versões*: Objetos entram em um sistema de informação digital ao serem criados digitalmente ou convertidos para o formato digital. Versões múltiplas de um mesmo objeto podem ser criadas para preservação, pesquisa e disseminação. Alguns metadados administrativos e descritivos podem ser incluídos pelo criador.
- *Organização*: Os objetos são automaticamente ou manualmente organizados dentro da estrutura do sistema de informação digital, e metadados adicionais podem ser criados através de identificação, catalogação e processos de indexação. A organização dos objetos de informação poderá ser refletida por metadados de catalogação associados aos objetos. Práticas de catalogação e identificação estão frequentemente relacionadas com o que está sendo arquivado, como serão acessados e com os recursos disponíveis para gerenciamento do arquivo.
- *Pesquisa e recuperação*: Objetos armazenados e distribuídos estão sujeitos à pesquisa e recuperação pelos usuários. O sistema de informação cria metadados que registram algoritmos de recuperação, transações de usuários e efetividade de sistema em armazenamento e recuperação.
- *Utilização e Atualização*: Objetos recuperados são utilizados, reproduzidos, e modificados. Metadados relacionados a anotações de usuários, registros de propriedade intelectual, e controle de versão podem ser criados.

- *Preservação e Descarte*: Para assegurar disponibilidade continuada, objetos de informação sofrem processos tais como atualização, migração de mídia de armazenamento, migração de *hardware* e *software*, e conferência de integridade. Aqueles que se tornarem inativos ou que não forem mais necessários podem ser descartados. Metadados podem documentar tanto atividades de descarte como de preservação, para que objetos possam continuar existindo independentemente do sistema de armazenamento e recuperação que atualmente estiverem sendo utilizados. Assim, metadados técnicos, descritivos e de preservação que documentem como cada objeto foi criado e mantido, seu comportamento, e como este se relaciona com outros objetos de informação, serão todos essenciais. É importante destacar que para objetos de informação permanecerem acessíveis e inteligíveis com o passar do tempo, também são necessárias a conservação e migração de seus metadados.

Metadados podem vir de diversas fontes. Podem ser providos por um humano (criador, profissional de TI ou usuário), criados automaticamente por programas de computador ou de outro tipo de equipamento (agentes), ou deduzidos através de uma relação com outro recurso, como um *hyperlink* [Marco2000].

A extração dos metadados pode se dar de forma manual ou automática [AMEGA2005]. Muitas características de baixo nível (cor, palavras-chave etc.) podem ser extraídas de modo completamente automático, enquanto que características de alto nível (informações semânticas elaboradas) normalmente necessitam da interação humana.

Atualmente equipamentos de produção de objetos digitais vêm crescentemente gerando informações descritivas de baixo nível de forma automática ou semi-automática [HM2002]. Por exemplo, máquinas fotográficas digitais que registram o momento e lugar de cada imagem capturada; filmadoras que registram título, produtor e mensagens de direitos autorais e permissões de uso. Tais metadados poderiam ser enriquecidos com informações de alto nível, como por exemplo: detalhes sobre cenas produzidos automaticamente via análise de fatos já registrados em conteúdos auditivos e visuais do fluxo de dados de vídeo.

Na criação de metadados voltados a ambientes digitais semânticos, há pelo menos três áreas desafiadoras a serem enfrentadas: descrição dos objetos de informação, geração de metadados efetivos e interoperabilidade. É justamente no contexto dessas três áreas de

desafios que esta Tese vem contribuir. Existem vários métodos de extração automática de metadados de documentos textuais na *Web*, como por exemplo: extração de palavras-chave, sumarização e categorização. Alguns desses métodos serão apresentados no Capítulo 4, e experimentos serão realizados.

O conteúdo dos metadados de um objeto de informação dependerá do tipo do objeto e do propósito de cada metadado, bem como do nível de uso destes (global, regional, comunidade etc.). É necessária flexibilidade para capturar informação útil para uma variedade de objetos de informação e propósitos.

Para redução de custos financeiros, de tempo, de erros de digitação e de mau emprego de vocabulário padrão, técnicas automatizadas para geração de metadados devem ser exploradas ao máximo possível [Hodge2000]. Sistemas de informação, em sua grande maioria, contêm um número elevado de processos manuais de geração de metadados que acabam por consumir demasiadamente o tempo de trabalho de sua equipe, tornando a solução não escalável e impossível de ser mantida ao longo do tempo.

À medida que projetistas de sistemas vão respondendo às necessidades de incorporar e gerenciar metadados em sistemas de informação, e se interessam também em solucionar problemas relativos à manutenção destes durante o tempo de vida dos objetos associados, muitos mecanismos adicionais automáticos para associar metadados aos seus objetos de informação provavelmente aparecerão.

3.5 Padronização de Metadados no Contexto da WEB

Apesar de muitos acreditarem que o conceito de metadados é novo no âmbito da Ciência da Computação, este já era conhecido antes da década de 70, quando surgiram os primeiros repositórios de metadados comerciais (denominados *dicionários de dados*). Estes, entretanto, centravam-se unicamente no dado, fornecendo informações como: definições, relacionamentos, origem, domínio, uso e formato. Na década de 90 surgiram os repositórios de metadados para ambiente cliente-servidor. A introdução de sistemas de suporte à decisão (DSS) veio exigir um maior esforço na busca de arquiteturas interoperáveis e flexíveis a mudanças, as quais puderam ser simplificadas com a utilização de metadados por profissionais experientes no uso efetivo destes.

Os esforços de padronização de metadados inicialmente focavam assuntos relacionados a intercâmbio, enfatizando o formato de arquivos exportados em lotes (*batch*). Com o crescimento da demanda por iniciativas baseadas em metadados, novos esforços de padronização foram criados, cada qual com sua filosofia e conceitos, sendo hoje representados por: fabricantes de *software*, consumidores da indústria (interessados em compartilhamento e intercâmbio de dados e metadados, dentro de suas próprias organizações) e os mantenedores de padrões internacionais (oriundos principalmente das organizações de pesquisa, e todos defendendo a criação de um mecanismo industrialmente neutro, tecnologicamente independente e padronizado) [Tannenb2001].

Para que metadados alcancem efetividade em suas diversas aplicações, maximizem sua longevidade e permitam processamento tão automatizado quanto possível, é essencial que sua estrutura, semântica e sintaxe estejam em conformidade com os padrões internacionais mais utilizados [Greenberg2003, Gilliland2000]. No âmbito da *Web*, o *Dublin Core Metadata Element Set* (DCMES), geralmente denominado apenas de *Dublin Core* (ou DC) [DublinCore], e o RDF são duas recomendações de dois organismos diferentes que foram apresentadas para solucionar o problema de descrição e recuperação de recursos eletrônicos na *Web*: respectivamente o DCMI – *Dublin Core Metadata Initiative*, e o W3C.

Para ser utilizado em larga escala, o DCMI optou por definir de forma ampla a semântica do DC, deixando as questões ligadas à sintaxe abertas e indefinidas. Já o RDF investiu nas regras sintáticas onde o DC pode ser embebido. Esta é a razão pela qual o DC e o RDF combinam tão bem. O DC não é um componente direto do RDF, porém é um dos primeiros esquemas testados com o RDF.

Outros padrões de elementos de metadados menos utilizados são os seguintes: LTSC (IMS) – *Learning Technology Standards Committee (Instructional Management System)* [LTSC-IMS], ADL (SCORM) - *Advanced Distributed Learning (Sharable Content Object Reference Model)* [ADL-SCORM], FGDC - *Content Standard for Digital Geospatial Metadata* [FGDC] e GILS - *Global Information Locator Service* [GILS].

Na maioria dos casos, o conjunto de descritores do *Dublin Core* é embutido no próprio objeto digital (HTML, XML e outros), entretanto, estes podem também ser armazenados em repositórios separados.

O conjunto de metadados descrito pelo DC é composto de 15 elementos, cujas características principais são a simplicidade na descrição dos recursos, entendimento semântico universal (dos elementos), escopo internacional e extensibilidade (o que permite sua adaptação às necessidades adicionais de descrição). O *Dublin Core* não tem a intenção de substituir modelos mais ricos como o código AACR2/MARC [MARC], mas apenas fornecer um conjunto básico de elementos de descrição que podem ser usados por catalogadores ou não catalogadores para simples descrição de recursos de informação.

Como já dito no capítulo anterior, apesar de oferecer uma estrutura representacional relativamente simples e aplicável para uma gama extensiva de recursos de domínio não específico, o DC é também uma ontologia porque impõe um modelo conceitual formalmente definido que facilita o processo automatizado necessário para dar suporte ao compartilhamento de conhecimento por sistemas.

O DC é um padrão adotado mundialmente e tem sido utilizado em vários projetos que buscam um entendimento entre diferentes comunidades de usuários. Os principais elementos de metadados do *Dublin Core* encontram-se listados na Tabela 3.3. As descrições em negrito correspondem aos valores adotados na atribuição de conteúdo aos metadados gerados automaticamente pelo Modelo SIGAMeta definido nesta Tese.

Tabela 3.3 – Elementos de Metadados do *Dublin Core* (<http://purl.pt/201/1/#dct1>)

<i>Elemento</i>	<i>Descrição</i>
Title	Um Título é o nome pelo qual o recurso é formalmente conhecido.
Subject	Tópicos do conteúdo do recurso. Tipicamente, um assunto deverá ser expresso por palavras-chave, frases (categoria) , ou códigos de classificação que descrevem o conteúdo do recurso. Recomenda-se a seleção de termos de vocabulários controlados, ou de sistemas de classificação formais.
Description	Uma breve descrição do trabalho. Pode incluir um resumo , um índice, uma referência a uma representação gráfica do conteúdo, uma descrição textual etc.
Author ou Creator	A entidade responsável pela existência do recurso: pessoa, organização ou serviço.
Publisher	A entidade responsável por tornar o recurso disponível na presente forma , tal como uma casa publicadora, um departamento de uma instituição, uma entidade, uma pessoa.
OtherAgent ou Contributor	Pessoas, além do(s) autor(es), que contribuíram substancialmente para o trabalho (por exemplo, artistas e editores, organização ou serviço).
Date	Uma data associada a um evento do ciclo de vida do recurso. Tipicamente, uma Data deve ser associada à criação ou disponibilidade do recurso. Recomenda-se para codificação de valores de datas um perfil da norma ISO 8601

	[http://purl.pt/201/1/#w3cdtf], segundo o formato AAAA-MM-DD .
ObjectType	A natureza ou gênero do conteúdo do recurso. Tipos incluem termos descrevendo categorias genéricas, funções, gêneros, ou níveis de agregação para o conteúdo. Recomenda-se a seleção de valores a partir de vocabulários controlados, como por exemplo, a lista do documento de trabalho <i>The DCMI Type Vocabulary</i> [http://dublincore.org/documents/dcmi-terms/#H5]), que recomenda: <i>Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text</i> .
Form ou Format	Descreve a manifestação física ou digital do recurso. O Formato deve incluir o tipo de meio do recurso, ou as suas dimensões. Este elemento deve ser usado para determinar <i>software, hardware</i> ou qualquer tipo de equipamento necessário para reproduzir ou operar o recurso. Exemplos de dimensões incluem tamanho e duração. Como boa prática recomenda-se a seleção de valores a partir de vocabulários controlados, como por exemplo a lista da <i>Internet Media Types</i> [http://www.mhaviila.com.br/topicos/web/http_mime.html] definindo formatos e meios: <i>text/html, text/css, text/xml, application/pdf, application/zip, image/jpeg, image/gif, audio/peg, video/apeec</i> etc.
Identifier	Uma referência não ambígua ao recurso. Recomenda-se o uso de uma cadeia de caracteres ou um número de acordo com um sistema de identificação formal , como por exemplo, o <i>Uniform Resource Identifier</i> (URI) (incluindo o <i>Uniform Resource Locator</i> (URL)), o <i>Digital Object Identifier</i> (DOI) e o <i>International Standard Book Number</i> (ISBN).
Relation	Uma referência a um recurso relacionado. Recomenda-se referir o recurso através de uma cadeia de caracteres ou número em conformidade com um sistema de identificação formal.
Source	Outras fontes (caso se aplique) das quais o trabalho se deriva. Recomenda-se a referência ao recurso fonte através de um identificador em conformidade com um sistema de identificação formal.
Language	O idioma do conteúdo intelectual do recurso. Como boa prática recomenda-se o uso do RFC 1766 [http://www.ietf.org/rfc/rfc1766.txt], o qual inclui um código de língua de duas letras (retirado da norma ISO 639 [http://www.oasis-open.org/cover/iso639a.html]), seguido opcionalmente por um código de duas letras para o país (retirado da norma ISO 3166 [http://www.oasis-open.org/cover/country3166.html])). Por exemplo, “pt” para Português, “br” para Brasil ou “ pt-br ” para o Português do Brasil.
Coverage	A área geográfica que o trabalho engloba, se aplicável. Inclui uma localização espacial (o nome de um lugar ou coordenadas geográficas), um período no tempo (a sua designação, data, ou intervalo de tempo), ou jurisdição (o nome de uma entidade administrativa). Como boa prática recomenda-se a seleção de valores de vocabulários controlados (como por exemplo o "Thesaurus of Geographic Names", devendo ainda ser usados, quando apropriado, preferencialmente nomes de lugares e designações de períodos no tempo, em vez de identificadores numéricos tais como coordenadas ou intervalos de datas.
Rights	Direitos ou outras propriedades intelectuais especificando as condições através das quais o trabalho pode ou não ser usado. Tipicamente, este elemento deverá conter uma declaração de gestão de direitos sobre o recurso, ou uma referência a um serviço que fornecerá essa informação . Tal poderá compreender informação sobre direitos de propriedade intelectual, direitos de autor ou outros. A ausência deste elemento não permite formular qualquer hipótese válida sobre quaisquer direitos sobre o recurso.

Descoberta de Conhecimento em Texto

Os capítulos anteriores tiveram a oportunidade de destacar a importância dos metadados na construção de ambientes digitais semânticos. Metadados, tendo o potencial de registrar conteúdo, contexto e estrutura de objetos de informação, têm oferecido meios efetivos para viabilizar não apenas atividades de descoberta de conhecimento, mas também diversas outras, como gerenciamento, preservação, disseminação e direitos autorais.

Como será apresentado neste capítulo, técnicas de Inteligência Artificial (IA) e de Recuperação de Informação (RI) vêm viabilizando a exploração de objetos de informação no sentido de extrair conhecimento de forma automática, podendo estas serem utilizadas na extração de metadados.

No contexto mundial atual, descobrir conhecimento é extremamente importante para tomada de decisões corretas e embasadas, a fim de se conquistar novos espaços nesse mundo globalizado. Entretanto, descobrir conhecimento não é uma tarefa fácil. Ela exige a análise, identificação e agregação de informações importantes a um conhecimento prévio, mudando seu estado inicial. Alguns dos grandes problemas na tomada de decisões estratégicas estão na grande quantidade de informação disponível, na variedade de fonte dessas informações e na forma de recuperá-las.

A possibilidade de simular em computadores o processo de descoberta de conhecimento viabilizou a manipulação dessa grande quantidade de informação, sendo esta área estudada hoje pela Inteligência Artificial e pela Recuperação de Informação. Entretanto, a análise automática de documentos é uma operação complexa que exige competência no plano lingüístico e no plano da ciência da computação, de forma que só é possível essa implementação por existir uma ligação íntima entre a lingüística e a gestão da informação, possibilitando a indexação baseada na análise semântica do conteúdo do documento.

Para as pessoas é relativamente fácil inferir conceitos a partir das palavras dos documentos, pois possuem um razoável conhecimento de gramática e do mundo que as cerca. Quando utilizando um sistema automático, o ideal é que a linguagem utilizada para

recuperar as informações esteja o mais próximo possível da linguagem natural, tornando fácil a formulação das expressões de busca. Entretanto, nem sempre os documentos disponíveis para pesquisa encontram-se no mesmo padrão lingüístico do usuário, obrigando os sistemas automáticos a terem a capacidade de inferir, a partir de palavras do texto, um conjunto de termos que seja representativo do conceito desejado, e que esteja dentro do padrão utilizado pelo usuário (que por sua vez também é muito variado). Daí a necessidade de um vocabulário controlado que sirva tanto aos usuários (para representar as expressões de busca) como aos sistemas de informação (para a representação do documento).

São inúmeros os fatores que determinam o grau de complexidade das idéias expressas nos textos redigidos em linguagem natural, como por exemplo: a diversidade de tipos de documentos, as características dos possíveis usuários e problemas léxico-semânticos (homonímia, polissemia, sinonímia etc.). Tudo isso impede a unicidade das palavras e uma comunicação sem ambigüidades. Em resumo, problemas morfológicos, sintáticos, fonológicos e semânticos afetam a forma, a estrutura, o significado e as funções do texto.

As primeiras aplicações de descoberta de conhecimento foram realizadas a partir de bancos de dados disponíveis em empresas, utilizando-se ferramentas com base em métodos estatísticos, métodos de Inteligência Artificial e métodos provenientes da área de Recuperação de Informação. Pouco a pouco esses métodos foram sendo adotados pela área de Sistemas de Informação, uma vez que suas coleções de dados poderiam ser fontes valiosas de conhecimento. Dessa forma, o objetivo deste capítulo é justamente apresentar tais tecnologias.

4.1 Recuperação de Informação

A adoção do conceito de “ambiente digital” pelas instituições voltadas à gestão e disseminação de informação, deu-lhes a responsabilidade de dar suporte a objetos de informação de diversas tipologias e complexidade. Nesse novo cenário, objetos digitais como textos, imagens, sons, vídeos e páginas *Web* requerem diferentes tipos de tratamento e representação, cada vez mais sofisticados, para viabilizar a descoberta de conhecimento de forma efetiva.

No âmbito da Ciência da Computação, Recuperação da Informação é uma área que estuda o armazenamento e recuperação de objetos de informação, de forma automática. Dado um conjunto de documentos (*corpus*), um Sistema de RI deve identificar aqueles que

atendam às necessidades de seus usuários, e apresentar os resultados de maneira que facilite sua exploração de forma amigável.

Os componentes de um processo de recuperação de informação podem ser representados conforme a Figura 4.1:

Figura 4.1 - Componentes de um Processo de Recuperação de Informação

Na Figura 4.1, *Documentos* representam objetos de informação que podem ser endereçados ou manipulados por um sistema de computação.

O processo de *Representação* objetiva descrever ou identificar cada documento, o que geralmente é realizado através do processo de *Indexação*. A indexação envolve a criação de estruturas de dados conceituais extraídos dos documentos, através da análise de seu conteúdo. Estas estruturas podem conter metadados e *tesauros*, bem como dados sobre características dos termos na coleção de documentos, como dados estatísticos. A representação de um documento pode inclusive ser utilizada como seu substituto, dentro de uma base de dados, para efeito de economia de armazenamento e velocidade de pesquisa.

O componente *Necessidades do Usuário* gera a *Especificação da Consulta*, que pode ser elaborada em linguagem natural ou artificial, fornecida pelo Sistema. Em geral, uma

consulta é composta de um conjunto de palavras que tentam exprimir a necessidade de informação do usuário. Este processo geralmente é uma tarefa difícil, visto que há freqüentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada. Essa distância é gerada pelo conhecimento limitado do usuário sobre o universo de pesquisa à sua disposição, e pelo formalismo das linguagens de consulta utilizadas por sistemas de recuperação.

Um método clássico de abordar este problema é considerar uma forma interativa de construção da consulta, onde o usuário formula uma consulta inicial, examina os resultados mediante suas necessidades e, se necessário, melhora a formulação da consulta. A principal vantagem deste método é que, após a primeira formulação, o usuário interage com o sistema abstraindo-se do processo de formulação, simplesmente identificando documentos como relevantes ou não. Outra vantagem é que o método provê um processo controlado de enfatizar alguns termos e diminuir a importância de outros [Cardoso2000].

O componente *Processo de Recuperação*, a partir da *Representação/Indexação* dos documentos e da *Especificação da Consulta* formulada, recupera uma lista de referências a documentos que supostamente seriam do interesse do usuário (*Documentos Recuperados*). Os índices construídos para a coleção de documentos são usados para acelerar esta tarefa. A lista de documentos recuperados é apresentada ao usuário segundo algum critério de classificação, como por exemplo, o grau de similaridade entre o documento e a especificação da consulta.

De acordo com os conceitos acima, em um primeiro momento pode-se concluir que é principalmente a má formulação da consulta em sistemas de recuperação que prejudica o seu desempenho, inundando os usuários com resultados irrelevantes. Observa-se, entretanto, que não basta apenas predizer nas consultas os termos utilizados na indexação dos documentos desejados, mas também evitar a recuperação de documentos irrelevantes (fora do contexto desejado), para que sejam minimizados os esforços dos usuários.

Como será mostrada nos experimentos realizados nesta Tese (Capítulo 6), diante de uma especificação de consulta, a identificação automática de documentos que poderiam ser considerados relevantes ou irrelevantes melhora se cada documento do universo de pesquisa explicitar a semântica de seu conteúdo. Em outras palavras, se os documentos estiverem representados e indexados de forma sofisticada. Isso leva a concluir que o componente

Representação e Indexação dos documentos também participa de forma fundamental no desempenho dos sistemas de recuperação.

Como já visto no Capítulo 2, a representação e a indexação de um documento podem ser sofisticadas o suficiente para explicitar a semântica de seu conteúdo, quando utilizando tecnologias e padrões que proporcionam automação, compartilhamento, integração e reuso de informação (Web Semântica). Explicitar a semântica de um documento significa declarar parâmetros que permitam medir a similaridade desse documento com os outros do universo de pesquisa, bem como com a especificação de consulta. Isto é o que caracteriza o problema da recuperação de informação efetiva: “dada uma especificação de consulta, como medir a proximidade (ou similaridade) desta especificação com o universo de documentos disponíveis, para se obter um conjunto de resultados mais próximo possível do ideal?”.

Segundo Aldenderfer e Blashfield [AB1984], a similaridade deveria ser medida de forma objetiva, reaplicável e quantitativa, ou seja, uma *métrica*. Porém nem sempre isto é possível devido à natureza das características dos objetos de informação, da área de aplicação, contextos das informações e perspectivas de uso.

No caso de documentos textuais, as características usadas para representá-los não são numéricas, ou seja, não são contínuas, pois os elementos utilizados são suas palavras. É importante dizer que nem todos os métodos de identificação de similaridade trabalham com atributos dessa natureza.

Há diversos modelos clássicos de recuperação de informação de natureza quantitativa, que apesar de terem sido criados nas décadas de 60 e 70, e aperfeiçoados nos anos 80, continuam ainda sendo utilizados pela maioria dos sistemas de recuperação, e por mecanismos de busca da *Web*. Além dos modelos quantitativos, modelos bem mais avançados e recentes têm sido propostos, que utilizam métodos de Inteligência Artificial e PLN. Estas são alternativas promissoras para serem utilizadas de forma ampla no futuro.

Nas próximas subseções, três modelos clássicos de recuperação de informação serão apresentados: Modelo booleano, Modelo booleano estendido e Modelo vetorial. Estes modelos consideram que cada documento pode ser descrito por um conjunto de termos de indexação considerado importante para descrever o conceito ou significado do documento (e conseqüentemente medir similaridades).

4.1.1 Modelo Booleano

No modelo booleano para recuperação de informação, os termos de indexação que representam um documento são comparados com a especificação da consulta, que consiste em um conjunto de termos combinados com operadores de Boole (AND, OR e NOT). Como o modelo é binário, cada documento ou corresponde à expressão booleana, ou não. Uma especificação de consulta do tipo $(t_1 \text{ OR } t_2 \text{ OR } \dots \text{ OR } t_n)$ recuperará documentos indexados por pelo menos um dos termos da especificação. Documentos indexados por todos os termos é tratado da mesma forma que um documento indexado por um único termo.

A Figura 4.2 ilustra termos de indexação (t_1 , t_2 e t_3) presentes (1) ou ausentes (0) em um conjunto de onze documentos (D) representados como uma matriz de vetores e como uma combinação booleana de termos.

Figura 4.2 - Matriz de Vetores e Combinação Booleana de Termos de Indexação

Apesar de bem formalizado e da facilidade de implementação, os problemas relacionados ao modelo booleano são bem conhecidos:

- Para buscas mais complexas, o usuário necessita estar familiarizado com a lógica booleana;
- A seleção dos termos para a consulta é difícil, especialmente se o domínio não é bem conhecido pelo usuário;

- A quantidade de documentos resultante não pode ser controlada, pois o usuário é incapaz de prever quantos elementos satisfarão à expressão booleana, podendo ser nenhum ou dezenas;
- O resultado da recuperação particiona o conjunto de documentos em dois subconjuntos: os que atendem à expressão booleana e os que não atendem. Presume-se então que não há nenhum mecanismo de ordenação dos documentos;
- Não há como atribuir graus de importância aos termos da expressão booleana, assumindo-se que todos possuem pesos iguais.

Apesar de suas limitações, vários mecanismos de busca são baseados neste modelo, como por exemplo, o *Alta Vista*, *Lycos* e *Excite*, na *Web*.

4.1.2 Modelo Vetorial

No modelo vetorial, os documentos são distribuídos em um espaço euclidiano multidimensional, onde cada ordenada corresponde a um termo de indexação (T_i). A similaridade é dada pela proximidade dos documentos nesse espaço, conforme pode ser visto na Figura 4.3.

Figura 4.3 – Distância Euclidiana

Cada documento (D_i) no espaço euclidiano é representado por um vetor de pesos associados aos termos de indexação daquele documento (T_i), que pode ser sua frequência ou uma função dela. Estes pesos traduzem a importância do termo para aquele documento. Da

mesma forma, uma especificação de consulta de usuário também pode ser representada por um vetor de pesos associados aos termos da consulta. A Figura 4.4 mostra a representação gráfica de uma especificação de consulta Q e dois documentos D_1 e D_2 , em um espaço vetorial formado por dois termos de indexação T_1 e T_2 .

Figura 4.4 – Representação de uma Especificação de Consulta e dois Documentos em um Espaço Vetorial Bidimensional

A similaridade entre dois documentos ou entre um documento e uma especificação de consulta é calculada através do co-seno do ângulo formado entre seus respectivos vetores de pesos [SM1983], desconsiderando-se seu tamanho. Dessa forma, na Figura 4.4, o co-seno do ângulo α_1 representa o grau de similaridade entre a especificação de consulta Q e o documento D_1 , e o co-seno do ângulo α_2 representa o grau de similaridade entre a especificação de consulta Q e o documento D_2 .

Para que se considere documentos de tamanhos diferentes (isto é, documentos com número de termos de indexação variados), os pesos devem estar normalizados, assumindo valores entre zero (0) e um (1). Pesos mais próximos de um representam termos de maior importância para a descrição do documento. Termos ausentes no documento recebem peso de valor zero.

Em um sistema real contendo um grande número de documentos e termos de indexação, uma matriz pode representar os vetores de pesos associados aos documentos (D_i) e à especificação de consulta (Q). Para um *corpus* com i documentos e t termos de indexação, a matriz de pesos (W_{dit}) pode ser representada da seguinte forma:

	T₁	T₂	...	T_t
D₁	W_{d11}	W_{d12}	...	W_{d1t}
D₂	W_{d21}	W_{d22}	...	W_{d2t}
·	·	·	·	·
D_i	W_{di1}	W_{di2}	...	W_{dit}
Q	W_{q1}	W_{q2}	...	W_{qt}

Figura 4.5 – Matriz de Pesos de *Corpus* contendo i Documentos e t Termos de Indexação

A similaridade do documento D_i com a especificação de consulta Q ($Sim(Q, D_i)$) é obtida através da fórmula F01 abaixo, quando os pesos dos termos de indexação já se encontram normalizados. Esta fórmula é chamada de co-seno ou produto interno normalizado [SB1988].

$$sim(Q, D_i) = \sum_{j=1}^t w_{qj} * w_{dij} \quad (\text{F01})$$

Nesta fórmula:

- $Sim(Q, D_i)$ representa a similaridade do documento D_i com a especificação de consulta Q ;
- W_{qj} representa o peso do termo j na especificação de consulta Q ;
- W_{dij} representa o peso do termo j no documento D_i ;
- t representa o número de termos.

Quando os pesos não se encontram normalizados, pode-se normalizá-los dentro da fórmula de similaridade F02 mostrada abaixo [SB1988]:

$$sim(Q, D_i) = \frac{\sum_{j=1}^t w_{qj} * w_{dij}}{\sqrt{\sum_{j=1}^t (w_{qj})^2 \sum_{j=1}^t (w_{dij})^2}} \quad (F02)$$

Para o espaço vetorial mostrado na Figura 4.4, onde o vetor da especificação de consulta Q e dos documentos D_1 e D_2 foram: $Q = (0.4, 0.8)$, $D_1 = (0.8, 0.3)$ e $D_2 = (0.2, 0.7)$, os cálculos de grau de similaridade pela fórmula F02 calculam um valor de 74% entre Q e D_1 , e de 98% entre Q e D_2 , conforme pode ser visto nos cálculos exibidos abaixo (foram considerados pesos não normalizados):

$$sim(Q, D1) = \frac{(0.4*0.8) + (0.8*0.3)}{\sqrt{[(0.4)^2 + (0.8)^2] \cdot [(0.8)^2 + (0.3)^2]}} = \frac{0.56}{\sqrt{0.58}} = 0.74$$

$$sim(Q, D2) = \frac{(0.4*0.2) + (0.8*0.7)}{\sqrt{[(0.4)^2 + (0.8)^2] \cdot [(0.2)^2 + (0.7)^2]}} = \frac{0.64}{\sqrt{0.42}} = 0.98$$

De forma análoga, pode-se calcular a similaridade dos documentos D_1 e D_2 , utilizando-se o mesmo procedimento acima.

A principal vantagem do modelo vetorial é a sua simplicidade de representar documentos em *corpus* textuais. Uma vez que operações com vetores podem ser executadas rapidamente, este modelo e suas variantes são freqüentemente utilizados no campo da RI [BR1999].

Uma vez que este modelo é baseado em palavras, ele traz problemas como sua alta dimensionalidade: o número de palavras diferentes em um *corpus* atinge facilmente centenas de milhares. Outro problema é a variação de estilos de escrita, palavras sinônimas, erros de grafia etc. Além disso, quaisquer duas palavras são consideradas, por definição, não-relacionadas.

Escolha dos termos de indexação: cálculo dos pesos [SB1988, SM1983, SSMB1997, CMNS1996]:

A escolha dos termos de indexação para um documento não é uma tarefa fácil, mesmo quando desempenhada por especialistas. Em geral, esta escolha é feita com base em propriedades estatísticas e/ou semânticas e sintáticas que facilitam a mensuração do potencial representativo do termo de indexação, como por exemplo, sua frequência.

O uso da frequência da palavra para cálculo do grau de importância parte do princípio de que as palavras não ocorrem aleatoriamente em um texto. Especialistas da área de lingüística afirmam que o escritor costuma usar e repetir os mesmos termos à medida que escreve, já que a inclusão de novos termos causaria a ele um esforço maior do que o necessário para explicar um assunto. Técnicas de Processamento de Linguagem Natural também são usadas, e algumas dessas serão comentadas mais adiante.

Existem várias formas de se atribuir pesos aos termos de indexação, como por exemplo: o esquema binário, a frequência absoluta do termo (tf), a frequência relativa e uma combinação da frequência absoluta com a frequência inversa ($tf*idf$).

O esquema binário usa os valores um (1) e zero (0) para revelar se um termo existe em um documento ou não, respectivamente, conforme visto no Modelo Booleano de recuperação de informação (Seção 4.1.1).

Para a frequência absoluta do termo, o peso corresponde à quantidade de ocorrência do termo em um documento. É normalmente denominado de tf (*Term Frequency*).

A frequência relativa de um termo consiste na normalização de sua frequência absoluta (tf). Isto é feito dividindo-se o número de ocorrências do termo pela frequência máxima de termo existente no documento. O resultado é um número entre 0 (zero) e 1 (um) que indica a importância relativa de uma palavra em um documento. Dessa forma, como já dito anteriormente, todos os documentos acabam por ter o “mesmo tamanho”, independente do seu número absoluto de palavras. Essa medida é também conhecida por frequência normalizada (*normalized term frequency*). Apesar de ser muito utilizada, tal medida não leva em conta os demais documentos de um *corpus*. Sua fórmula é a seguinte:

$$W_{ik} = \frac{tf_{ik}}{tf_{i\max}} \quad (\text{F03})$$

Nesta fórmula:

- W_{ik} representa o peso do termo k no documento i ;
- tf_{ik} representa a frequência do termo k no documento i ;
- $tf_{i\max}$ representa a frequência máxima de termo no documento i .

O esquema $tf*idf$ (*Term Frequency – Inverse Document Frequency*) é uma técnica que faz o balanceamento entre as características do documento (medida local - tf), e a importância de um termo no *corpus* (medida global - idf). Dessa forma, se um *corpus* possui N documentos e n_k é a quantidade de documentos que possuem o termo T_k , então o peso do termo k para o documento i é dado pela fórmula F04 abaixo:

$$w_{ik} = tf_{ik} * idf_k \quad (\text{F04})$$

Nesta fórmula:

- W_{ik} representa o peso do termo k no documento i ;
- tf_{ik} representa a frequência do termo k no documento i ;
- idf_k representa a frequência inversa do termo k no *corpus* C .

A frequência inversa de um termo (idf) tenta expressar a importância de uma palavra dentro de uma coleção. Quanto mais raro o termo dentro de um *corpus*, mais alto será o seu idf . Sua fórmula é a seguinte:

$$idf_k = \log(N / n_k) \quad (\text{F05})$$

Nesta fórmula, N é a quantidade de documentos no *corpus* C , e n_k é a quantidade de documentos que possuem o termo T_k . Assim, termos que ocorrem em muitos documentos são penalizados por oferecerem pouca discriminação.

Além dessas medidas, existem várias outras que também identificam a importância dos termos em um documento. Algumas delas são variações das anteriores. Outras, porém, são mais sofisticadas e mais difíceis de serem calculadas por necessitarem de informações relativas à distribuição de frequência das palavras em todo o *corpus*. Como exemplo, a literatura fornece a seguinte fórmula:

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^t (tf_{ik})^2 [\log(N / n_k)]^2}} \quad (\text{F06})$$

Nesta fórmula:

- w_{ik} representa o peso do termo k no documento i ;
- tf_{ik} representa a frequência do termo k no documento i ;
- N é a quantidade de documentos no *corpus* C ;
- n_k é a quantidade de documentos que possuem o termo T_k ;
- t é o número de termos no documento.

Todas essas medidas foram criadas para especialmente tratar de aplicações de categorização de textos, recuperação de informação, sumarização e na análise e compreensão de conglomerados. Nessa categoria, também se enquadram o *coeficiente de correlação de Schutze*, o *coeficiente de correlação de Hwee Ng*, o *algoritmo de termo-seleção*, e a *importância do termo* [SB1988, Lopes2004].

Uma vez calculados os pesos dos termos, é necessário então selecionar aqueles mais importantes, através de algum critério de corte. O corte pode ser realizado tanto a partir do próprio peso, estipulando-se um valor mínimo, como também pelo número de palavras, ou ainda, por ambas as formas.

4.1.3 Modelo Booleano Estendido

Este modelo foi proposto por Salton, Fox e Wu [SFW1983], e tenta unir a vantagem do formalismo das expressões booleanas com a precisão do Modelo Vetorial. Conforme visto no Modelo Booleano (Seção 4.1.1), uma especificação de consulta composta por onze

termos de indexação conectados por operadores AND (t_1 AND t_2 AND t_{11}), tratará da mesma forma documentos indexados por 10 termos ou documentos indexado por um único termo, excluindo-os. Entretanto, é de se esperar que muitos desses documentos excluídos provavelmente possuiriam informações importantes e aproveitáveis pelos usuários.

Pela proposta do Modelo Booleano Estendido, flexibiliza-se o Modelo Booleano através da introdução do conceito de relevância, permitindo a ordenação dos documentos recuperados, e, por outro lado, permite um maior poder às buscas, oferecido pelo Modelo Vetorial, através do uso de operadores booleanos.

Semelhante ao Modelo Vetorial, documentos e especificação de consulta são representados por vetores de pesos de seus respectivos termos de indexação, definindo seus posicionamentos num espaço vetorial. Por exemplo, considerar uma especificação de consulta Q composta de dois termos de indexação, $k1$ e $k2$. Uma representação espacial desses termos contém dois pontos extremos: ponto (1, 1), correspondendo a documentos que contêm ambos os termos $k1$ e $k2$; ponto (0, 0), correspondendo a documentos que não contêm nenhum dos termos. Nesta representação, a Distância Euclidiana máxima ou deslocamento máximo (d_{max}) entre os dois pontos é de 1,41, conforme pode ser visto na Figura 4.6.

$$d_{max} = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2} = 1.41$$

Figura 4.6 - Distância Euclidiana Máxima em Espaço Bidimensional

Dessa forma, para uma especificação de consulta $Q = k1 \text{ OR } k2$, a Distância Euclidiana de um documento nos pontos (w_{k1}, w_{k2}) deve ser $d \leq 1,41$, conforme Figura 4.7:

$$d_{\text{OR}} = \sqrt{(w_{k1} - 0)^2 + (w_{k2} - 0)^2} = \sqrt{w_{k1}^2 + w_{k2}^2}$$

Figure 4.7 - Distância Euclidiana em Espaço Bidimensional (operador OR)

Com base nas explicações acima, pode-se considerar a variação da distância da origem (ponto (0,0)) até o ponto (w_{k1}, w_{k2}) como sendo medidas de semelhança (similaridade) de documentos, onde aqueles mais próximos da origem seriam menos relevantes que aqueles mais próximos do ponto (w_{k1}, w_{k2}) .

Para o caso de uma especificação de consulta $Q = k1 \text{ AND } k2$, a medida de similaridade pode ser computada através da subtração da Distância Euclidiana máxima entre os pontos (1, 1) e (w_{k1}, w_{k2}) conforme Figura 4.8:

$$d_{\text{AND}} = \sqrt{2} - \sqrt{(1 - w_{k1})^2 + (1 - w_{k2})^2}$$

Figure 4.8 - Distância Euclidiana em Espaço Bidimensional (operador AND)

Para comparar medidas de similaridade em documentos de tamanhos variados, é preciso que se normalizem todas as distâncias, dividindo-as pelo valor da distância máxima (d_{max}). Assim, para especificações de consulta OR e AND, obtêm-se as seguintes fórmulas, em um espaço bidimensional (Figura 4.9):

$$Sim(Q_{k1 \text{ OR } k2}, D) = \sqrt{\frac{w_{k1}^2 + w_{k2}^2}{2}}$$

$$Sim(Q_{k1 \text{ AND } k2}, D) = 1 - \sqrt{\frac{(1 - w_{k1})^2 + (1 - w_{k2})^2}{2}}$$

Figura 4.9 – Medida de Similaridade Normalizada (operadores OR e AND)

Dessa forma, considerando agora qualquer número de termos de indexação (m), podem-se obter expressões que dependem de um parâmetro p , utilizando-se o conceito matemático de norma. O valor desse parâmetro é definido durante a formulação da especificação de consulta. Assim, obtêm-se as fórmulas mostradas abaixo:

$$Sim(Q_{k1 \text{ OR } k2}, D) = \left(\frac{w_{k1}^p + w_{k2}^p + \dots + w_{km}^p}{m} \right)^{1/p}$$

$$Sim(Q_{k1 \text{ AND } k2}, D) = 1 - \left(\frac{(1 - w_{k1})^p + (1 - w_{k2})^p + \dots + (1 - w_{km})^p}{m} \right)^{1/p}$$

Figura 4.10 – Medida de Similaridade Normalizada para m Termos (OR e AND)

O cálculo de similaridade para uma especificação de consulta composta de operadores AND e OR é realizado através de cálculos de partes da expressão.

Finalmente, para se incluir os pesos da especificação de consulta, multiplicam-se estes pelos pesos dos termos de indexação do documento. Assim, para uma especificação de

consulta que consiste em apenas dois termos, o cálculo de similaridade é dado pelas seguintes fórmulas:

$$Sim(Q_{k1 \text{ OR } k2}, D) = \left(\frac{q_{k1}^p w_{k1}^p + q_{k2}^p w_{k2}^p}{q_{k1}^p + q_{k2}^p} \right)^{1/p}$$

$$Sim(Q_{k1 \text{ AND } k2}, D) = 1 - \left(\frac{q_{k1}^p (1 - w_{k1})^p + q_{k2}^p (1 - w_{k2})^p}{q_{k1}^p + q_{k2}^p} \right)^{1/p}$$

Figura 4.11 – Medida de Similaridade com Pesos da Especificação de Consulta (OR e AND)

A norma p coloca propriedades interessantes [Garcia2005]:

- Quando $p = 1$, um espaço vetorial simples é obtido. A distinção entre OR e AND desaparece e a similaridade entre a especificação de consulta e documentos pode ser computada pelo produto interno entre seus respectivos pesos, isto é, o resultado é semelhante ao obtido no modelo vetorial;
- Quando $p = 2$, este é equivalente à distância euclidiana normalizada;
- Quando p está entre infinito e 1, os resultados produzidos são intermediários entre a busca booleana pura e busca do modelo vetorial;
- Quando $p = \text{infinito}$ (ou bastante alto), especificações de consulta AND e OR se comportam como as expressões booleanas convencionais, onde pesos não são incluídos. Além disso, $p = \text{infinito}$ resulta em uma similaridade cujo valor será o mínimo (AND) ou máximo (OR) do conjunto dos pesos do documento.

4.2 Recuperação de Informação na Web

A recuperação de informação na *Web* é geralmente realizada por mecanismos de busca, conhecidos também como *ferramenta de busca*, *site de busca*, *portais*, *engenho de busca* etc. Nesta Tese, o termo “mecanismo de busca” será utilizado para se referir a

sistemas de busca baseados na indexação automática de páginas *Web*, como também para aqueles que se baseiam na indexação manual (diretórios).

Um mecanismo de busca é definido como um conjunto constituído de computadores, índices, bases de dados e algoritmos, com o objetivo de analisar e indexar páginas *Web*. Os resultados dessa análise são armazenados numa base de dados, a fim de recuperar os endereços das páginas (URLs) que preenchem os requisitos indicados pelos usuários, no momento de uma consulta.

Considerando o tamanho atual da *Web* e sua característica de crescimento exponencial, é natural a indexação prévia de suas páginas, que inclusive devem permanecer em constante processo de atualização. Dessa forma, ao se realizar uma pesquisa na *Web*, o mecanismo de busca não consulta diretamente a *Web*, e sim a sua base de dados previamente construída e localizada em algum *site*. Outra informação importante é que nenhum mecanismo de busca tem em sua base de dados toda a *Web* catalogada. A indexação das páginas pelos mecanismos de busca pode ser realizada de forma manual ou automática.

Na indexação manual, a análise das páginas *Web* e a criação da base de dados são realizadas por profissionais, que também especificam uma hierarquia de assuntos e indexam as páginas segundo esta hierarquia (categorias e sub-categorias, setores de atividade econômica ou ramos do conhecimento). Para a recuperação da informação, o usuário caminha das categorias principais para as sub-categorias que estiverem mais de acordo com a classe de informação desejada, até chegar ao nível de especificidade procurado. Nesta modalidade, não existem programas que rastreiam a *Web* em busca de páginas a serem catalogadas, ficando ao encargo de autores de páginas a iniciativa de solicitar o cadastramento destas, repassando também a sua categorização e resumo do assunto tratado.

Esse tipo de mecanismo de busca pode operar ou não com a presença de uma equipe de editores que aprovará ou rejeitará a inclusão do *site* a ser inscrito, julgando com base nos critérios de relevância adotados. São exemplos desse tipo de mecanismo de busca o *Yahoo!* (<http://br.yahoo.com/>), *LookSmart* (<http://search.looksmart.com/>), *Open Directory* (<http://www.dmoz.org/>), *Cadê?* (<http://www.cade.com.br/>) e *Snap* (<http://www.snap.com/>).

Como consequência de suas próprias características, mecanismos de busca de indexação manual geram uma cobertura incompleta e irregular da *Web*. Por outro lado, a probabilidade de se obter resultados precisos em buscas de assuntos que se enquadram nas categorias oferecidas é muito alta, sendo, portanto, uma ótima opção para começar uma pesquisa na *Web*. Além do mais, quando se utiliza o julgamento humano na escolha das páginas a serem catalogadas, dificilmente se encontrarão páginas de conteúdo sem qualidade. Quanto a seu modo de pesquisa, a indexação manual permite tanto o uso de palavras-chave como também a navegação em sua hierarquia de categorias.

Com relação aos mecanismos de busca que utilizam indexação automática, estes possuem três componentes principais. Primeiro, um programa de computador (*robôs*, *spiders*, *crawlers*, *worm*) que percorre os *sites* da *Web*, selecionando URLs de páginas e algumas informações de conteúdo para verificar se estas devem ser indexadas em sua base de dados. Os critérios usados para seleção dificilmente são disponibilizados pelas empresas. Há ainda mecanismos que copiam não apenas a URL das páginas selecionadas, mas também todo o seu conteúdo.

O segundo componente é a base de dados de URLs selecionadas, que também pode ser chamada de índice ou catálogo. Essas informações são indexadas automaticamente através de métodos também não revelados pelas empresas. Entretanto, em geral utilizam-se técnicas estatísticas (como as vistas no Modelo Vetorial) e de Processamento de Linguagem Natural (PLN) para atribuir pesos às palavras. PLN possui técnicas como a de extração de radicais (*stemming*) para normalizar os termos de indexação e exclusão de palavras que não carregam contexto, como pronomes, preposições, artigos e conjunções (*stopwords*). Estas técnicas serão discutidas na próxima seção.

O terceiro componente é o programa de busca propriamente dito, que é acionado cada vez que um usuário realiza uma pesquisa. Nesse momento, o programa percorre a base de dados em busca dos endereços das páginas que contenham palavras, expressões ou frases informadas na consulta, retornando ao usuário os endereços encontrados. Exemplos de mecanismos de busca que utilizam indexação automática são: *Google* (<http://www.google.com.br/>), *Alta Vista* (<http://www.altavista.com/>), *All the Web* (<http://www.alltheweb.com/>) e *Excite* (<http://www.excite.com/>).

Existe ainda um terceiro tipo de mecanismo de busca, denominado meta-mecanismo, que se utiliza de diversos outros mecanismos de busca, com o objetivo de obter uma maior cobertura das informações da *Web*. Estes são programas mais simples, pois têm apenas a função de repassar aos mecanismos de busca a expressão de consulta do usuário, agregar as listas de resultados, eliminar duplicações e apresentar o resultado final. Um exemplo desse mecanismo é o *MetaCrawler* (www.metacrawler.com/).

Apesar dos diversos mecanismos de busca oferecidos pela *Web*, dificilmente se obtém resultados precisos freqüentes. Como já dito anteriormente, um dos fortes motivos dessa baixa precisão está na ausência de elementos que explicitem a semântica das informações da *Web*, como os metadados, dificultando os mecanismos de busca na seleção das páginas e na indexação efetiva de suas bases de dados. Técnicas de Inteligência Artificial propõem algumas soluções na extração de conhecimento dos objetos de informação, e as principais serão apresentadas na próxima seção.

4.3 Descoberta de Conhecimento em Texto

A Inteligência Artificial busca compreender o conhecimento e o processo da inteligência humana, desenvolvendo formalismos ou mecanismos de representação, armazenamento e utilização de conhecimento, através de uma linguagem artificial, tentando aproximar essa representação o melhor possível da forma natural realizada pelo homem. Dentro da IA, *Descoberta de Conhecimento* pode ser classificada em duas sub-áreas: Descoberta de Conhecimento em Dados Estruturados (KDD) e Descoberta de Conhecimento em Textos (KDT), também conhecida como Mineração de Texto (*Text Mining*).

O processo de KDT engloba técnicas e ferramentas automáticas inteligentes que auxiliam na análise de grandes volumes de dados, com o objetivo de extrair padrões ou conhecimento interessantes e não triviais, beneficiando não apenas usuários de documentos da *Web*, mas também de qualquer domínio que utiliza textos não estruturados. Suas principais técnicas são:

- Análise de conglomerados (*clustering*): descoberta automática de agrupamentos através da análise de afinidades ou similaridades;

- Categorização: identificação dos tópicos principais de um documento, associando-os a uma ou mais categorias pré-definidas;
- Extração de informação (EI): procura encontrar informações específicas dentro dos textos, como por exemplo, nomes próprios e de companhias. A EI divide o documento em partes relevantes e irrelevantes, e preenche modelos pré-definidos com a informação extraída. Esses modelos são utilizados para agilizar a RI;
- Sumarização: é a sintetização de um documento, porém sem perder a sua idéia principal.

Basicamente, as etapas mais importantes num processo de KDT são as seguintes:

- Seleção de documentos relevantes à análise: a partir de uma base de documentos textuais de qualquer origem (*Web*, CD-ROM), localizar e recuperar documentos relevantes, objetivando a redução do espaço de investigação;
- Limpeza ou padronização dos dados: uma vez selecionados os documentos para análise, estes deverão passar por um processo de eliminação de erros ortográficos, palavras que não carregam contexto (*stopwords*), e variações morfológicas. Técnicas de PLN auxiliam nesta tarefa, e serão vistas mais adiante;
- Seleção de dados relevantes para análise: após o processo de limpeza ou padronização, as palavras dos documentos devem passar por alguma técnica de aferição de relevância (pesagem), que geralmente envolve filtragem sintática/semântica e/ou estatística. Algumas dessas técnicas já foram vistas na Seção 4.1;
- Aplicação de algoritmo de mineração: uma vez selecionados os dados relevantes de cada documento, pode-se aplicar a técnica ou método de mineração desejado (conglomerados, classificação, categorização, extração, sumarização etc.) para se identificar padrões e relacionamentos entre os documentos.

Técnicas para extração de padrões e conhecimento de bases de dados textuais normalmente se baseiam nas palavras nelas contidas. Entretanto, nem todas as palavras que aparecem em um documento são importantes para representá-lo. Como visto na seção 4.1, há vários processos estatísticos e quantitativos que objetivam selecionar palavras que seriam consideradas mais representativas dentro de um documento. O problema é que normalmente essa escolha é baseada na frequência das palavras nos documentos ou no *corpus*, e palavras

soltas podem não representar adequadamente o conteúdo de um documento, por se encontrarem descontextualizadas.

Métodos utilizados em PLN vêm impulsionar a extração de padrões e conhecimento em textos. O PLN é um conjunto de técnicas computacionais para análise de textos em um ou mais níveis lingüísticos, com o propósito de simular o processamento humano da língua. Comparados com os métodos estatísticos, PLN são mais complexos e exigem maior poder e recursos computacionais, além de necessitarem de conhecimento de apoio que indique as gramáticas e heurísticas que regem a língua utilizada nos textos. Quando utilizados em conjunto com métodos estatísticos, métodos de PLN proporcionam resultados interessantes na escolha de termos representativos de documentos [LJ1996]. Como será visto no Capítulo 6, os experimntos dessa Tese utilizaram a combinação desses métodos, obtendo resultados importantes. Abaixo são descritos alguns desses métodos.

O *stemmer* é uma técnica de padronização de vocabulário (logo, dependente do idioma), que busca identificar termos similares e reduzi-los a um único radical, eliminando suas diferenças morfológicas ou lexicas (flexões e variações da palavra). Uma das técnicas mais conhecidas é a de Porter [Porter1980], que parte da hipótese de que termos que possuem a mesma raiz (radical) costumam ter significados similares. É um algoritmo baseado em regras heurísticas de remoção de sufixos (*suffix stripping*), sendo amplamente utilizado devido à sua simplicidade e eficiência. Uma versão para textos em português está disponível em (<http://snowball.tartarus.org/portuguese/stemmer.html>), incluindo toda a sua documentação.

Um outro algoritmo é o *Stemmer Portuguese*, também conhecido por RSLP (Removedor de Sufixos para a Língua Portuguesa) [Orengo2004, OH2001], cujo fluxograma é apresentado a seguir na Figura 4.12. Implementado originalmente em Linguagem C, o algoritmo é composto de oito passos (representados pelos retângulos na Figura 4.12) que devem ser executados seguindo uma ordem definida. Cada passo possui um conjunto de regras, porém nem todas as regras são necessariamente aplicadas.

Figura 4.12 – Fluxograma do Algoritmo *Stemmer Portuguese* (RSLP)

As principais diferenças entre os dois *stemmers* estão relacionadas com questões de morfologia e de tratamento dos caracteres de acentuação. A versão em português do algoritmo de Porter é uma tradução da versão original inglesa, não tratando, portanto, de problemas como formas aumentativas e diminutivas e do gênero feminino, já que isto não ocorre no inglês. Já o algoritmo RSLP foi projetado exclusivamente para a língua portuguesa, tratando assim de assuntos específicos de sua morfologia. Com relação a caracteres de acentuação, a versão portuguesa de Porter usa uma representação intermediária para todos os caracteres de acento (por exemplo, a vogal “ã” é tratada como uma vogal seguida de uma consoante “a~”). O *stemmer* RSLP trata todos os acentos como eles são. Nos experimentos dessa Tese, esses dois algoritmos de *stemmer* foram utilizados e avaliados.

Um outro método utilizado em PLN é o uso de dicionário ou *thesaurus*. Um *thesaurus* pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos para ajudar os usuários a encontrar a informação

de que eles precisam. Uma vez que usuários diferentes definem a mesma *query* utilizando termos diferentes (sinônimos, abreviações, acrônimos, ortografias alternativas etc.), um *thesaurus* pode ser usado para mapear tais termos naqueles que foram usados na indexação dos documentos.

Outra estratégia em PLN é considerar não apenas termos isolados, mas também aqueles compostos (duas ou mais palavras juntas que co-ocorrem), por se considerar que palavras compostas podem carregar uma maior semântica que àquelas isoladas. Para um melhor resultado na seleção de palavras compostas representativas, é importante combinar a co-ocorrência com métodos estatísticos (frequência).

A remoção de *stopwords* também é importante na seleção de palavras representativas. *Stopwords* são palavras muito frequentes dentro de uma coleção de documentos, sendo, portanto, inúteis para representar e distinguir os documentos uns dos outros. Por exemplo, preposições, pronomes, artigos, conjunções e outras classes de palavras auxiliares na construção sintática das orações. Também podem ser palavras que carregam contexto mas que apresentam uma incidência muito alta dentro de um domínio.

A utilização mais importante do PLN está na interpretação do conteúdo dos documentos, a fim de gerar uma representação adequada destes. No entanto, PLN não elimina a necessidade da utilização de métodos estatísticos e deve ser vista como uma ferramenta complementar.

Uma vez que os experimentos dessa Tese envolvem as técnicas de KDT de categorização e sumarização, nas próximas duas seções estas serão apresentadas.

4.4 Categorização de Textos

A categorização de textos é uma técnica de KDT que classifica documentos de acordo com categorias de conhecimento pré-definidas, com o objetivo maior de organizar uma base de dados textual. Estas categorias são representadas por termos (ou termo) que traduzem o significado do conceito associado.

Classificar não é apenas uma tendência natural do conhecimento, na medida em que é imprescindível acompanhar o crescimento da informação disponível; é igualmente um

imperativo da inteligência humana que tem a necessidade de organizar informações em sua mente, para poder coordená-las. A técnica de categorização, aplicada à disseminação, recuperação de informação e navegação na estrutura do conhecimento, resulta num maior valor competitivo para as organizações, visto que impulsiona o acesso efetivo à informação pretendida pelo usuário. Com a categorização, é possível, por exemplo, filtrar e encaminhar mensagens de correio eletrônico, organizar notícias, resumos e arquivos de publicações periódicas.

Na gestão de documentos textuais, a principal aplicação da categorização dá-se na indexação das informações, cuja preocupação é a descrição do conteúdo dos documentos, extraindo-se as palavras ou expressões mais significativas e vinculando-as a categorias. Estas categorias são ordenadas de forma a constituírem índices que serão utilizados em buscas e descobertas. Uma vez identificado o conteúdo de um documento, ele é armazenado na respectiva categoria. O processo de recuperação dá-se pela repescagem de todos os documentos pertencentes à determinada categoria.

Na navegação na estrutura do conhecimento (*browsing*), a categorização consiste em organizar os assuntos numa estrutura hierárquica, com categorias e subcategorias vinculadas ao texto pelo conteúdo, o que possibilita ao usuário navegar por essa hierarquia até encontrar a informação que precisa. A navegação compreende duas etapas fundamentais: a busca e a classificação. Na busca, o usuário acessa *links* e *sublinks* para pesquisar algum tema de interesse; na classificação, um documento é catalogado em determinada classe, depois de passar por um processo de triagem que definiu a sua pertinência e a sua relação com as possíveis categorias.

Como visto na seção 4.2, o funcionamento de alguns mecanismos de busca da *Web* (os diretórios) é dependente da existência de uma prévia categorização da informação, que ainda é quase inteiramente realizada por intervenção humana, sendo, portanto, muito cara.

Como já explicitado no início desse Capítulo, a riqueza dos matizes semânticos da linguagem impõe algumas limitações na automatização da exploração do conteúdo de documentos, dificultando uma recuperação perfeita da informação textual. Mas estas limitações não travaram o desenvolvimento de estudos e técnicas para aperfeiçoar a

eficiência da categorização de textos, e técnicas de PLN, KDT, RI e métodos estatísticos, já vistos nas seções anteriores, têm sido utilizadas com sucesso.

Basicamente existem duas formas de gerar categorias para documentos textuais. A primeira é a criação de um *thesaurus* que define o conjunto de termos específicos para cada domínio e a relação entre eles. Uma segunda forma, e a que será adotada nesta Tese, é conhecida como *Categorização assistida*. Este processo (ver Figura 4.13) consiste de duas etapas: geração das categorias (etapa de treinamento) e a categorização dos documentos. Esta abordagem oferece a vantagem de evitar o processo trabalhoso de construção de um *thesaurus*.

Figura 4.13 – Componentes de um Processo Automático de Categorização Assistida de Textos

O *corpus* de treinamento consiste num conjunto de textos (estático ou dinâmico), cuja categoria já é conhecida, e que é tido como um bom representante da categoria. Este *corpus* é submetido a um pré-processamento, onde operações de normalização (exclusão de caracteres especiais, figuras etc.), remoção de *stopwords*, substituição de sinônimos, processo de *stemmer* e outros podem ser aplicados aos textos. Após essa primeira filtragem, segue-se então a geração das características da categoria, que consiste em identificar os

termos que melhor representam o conteúdo dos textos, aplicando-se métodos estatísticos para cálculo de pesos (como aqueles já vistos na Seção 4.1), ou então métodos que consideram informações estruturais ou sintáticas das palavras.

Para informações estruturais, palavras que aparecem em títulos, subtítulos e outras marcações são consideradas mais importantes que as outras. Porém, identificar a estrutura de um documento não é tarefa muito simples e, a não ser que a estrutura do documento seja pré-determinada ou conhecida (como em HTML e XML), essa técnica não é muito viável. Com métodos sintáticos, técnicas de PLN são utilizadas para identificar a categoria sintática das palavras.

Para cada *corpus* de categoria diferente, esse processo de geração de características deve ser executado para gerar os diversos vetores que representam as categorias. Idealmente, a dimensão desses vetores seria infinita. Como isso não é possível, o desafio aqui está na sua redução baseada em algum critério de corte que não implique em perda de informação que seja relevante para a categorização, objetivando a redução de esforço computacional e a seleção das características significativas para a tarefa de categorização.

Na segunda fase (categorização de documentos), estes devem passar por métodos que meçam a sua similaridade com as diversas categorias disponíveis (gerada na primeira fase). Para tanto, é preciso que estes documentos também passem por um pré-processamento, semelhante ao descrito acima, para filtragem de termos não representativos. Como já visto na Seção 4.1.2, existem vários algoritmos de similaridade voltados à categorização de documentos.

4.5 A Sumarização de Textos

O crescimento das tecnologias de comunicação e disseminação de informação, em suas várias formas, tem trazido ao homem dificuldades de absorver, em tempo hábil, o que de fato é importante, neste grande volume de informação disponível. Aprender o que é essencial exige tempo, capacidade de identificar o que é relevante, e capacidade de mentalizar, de forma coerente, o conteúdo essencial. Estas questões remetem à necessidade de *sumarizar* as informações. O sumário é um resumo que tem o objetivo de passar ao leitor a idéia principal

de um texto, em poucas linhas. Podem ser utilizados na classificação de documentos bibliográficos e podem auxiliar atividade de recuperação de informação.

Dos vários tipos de sumários identificados ([BB1975, Cremmins1996, Sparck1999, HL1999]), quanto ao seu conteúdo os sumários podem ser:

- *Indicativos* - provêm uma idéia aproximada do que o texto informa, não podendo substituir o texto-fonte. Servem de referência para indicar se este merece uma leitura detalhada;
- *Informativos* - provêm uma versão resumida das partes importantes do conteúdo, sendo considerados substitutos do texto;
- *Críticos* – além de incluir toda a informação relevante, apresentam uma análise crítica do conteúdo do texto fonte.

Os *sumários indexadores* são semelhantes aos sumários indicativos, podendo ainda se apresentar como uma lista de itens. Quanto à audiência, os sumários podem ser:

- *Genéricos* – dirigem-se a um grupo amplo e não específico de usuário, sem considerar qualquer exigência de uso (sumário de propósito geral);
- *Específicos* - endereçam interesses específicos de grupo, ou mesmo exigências de uso individualizadas (sumário de propósito especial). Podem ser informações captadas através de termos de *queries*, perfis de interesses, informações de uso e outros.

Com relação à forma que são obtidos, os sumários são chamados de *Extratos* (sumários extrativos) quando são criados pela composição de seguimentos do texto fonte (orações, cláusulas, frases etc.). Se forem criados pela reformulação do conteúdo extraído, são ditos *abstratos*.

A *extração* é o processo de identificar seguimentos importantes dentro do texto; *abstração* é o processo de reformular os seguimentos com novos termos; *fusão* é o processo de combinar porções extraídas do texto fonte; e *compressão* é o processo de extrair partes sem importância do texto fonte. Em todos estes processos, é necessário que o sumário gerado seja coerente e esteja gramaticalmente correto.

Todas essas variações de sumário decorrem das diversas perspectivas, em função dos diferentes perfis de usuários e de suas tarefas desempenhadas.

4.5.1 A Sumarização Automática

Não há dúvidas que a tarefa de escrita de um texto resumido, com o objetivo de transmitir ou comunicar somente o que é importante (o sumário), é uma perspectiva fundamental no mundo moderno e globalizado. Entretanto, até que ponto modelos automáticos podem incorporar as principais características da tarefa de sumarização de textos, quando esta é realizada por humanos?

Resumidamente, a sumarização humana consiste em identificar segmentos relevantes de um texto e compô-los de forma coerente, para produzir os sumários correspondentes. Este é o objetivo da sumarização automática (SA): realizar a produção automática de sumários textuais e garantir correspondência entre os resultados automáticos e os textos fonte, de forma consistente e coesa (em termos gramaticais e de conteúdo). Embora pesquisadores, há algum tempo, venham investigando e disponibilizando ferramentas e métodos que automaticamente extraem ou abstraem conteúdo de fontes de informação, com o volume crescente da informação *on-line*, cada vez se torna mais difícil a geração de resumos significantes e oportunos.

Segundo Rino e Pardo [RP2003], o Projeto e Desenvolvimento (P&D) de um sistema automático de sumarização englobam, em geral, duas áreas importantes:

- a) Modelagem de procedimentos automáticos para escolha das informações e estruturação dos sumários a serem gerados:
 - Existem diversos modelos que envolvem o conhecimento lingüístico ou a manipulação estatística ou empírica das informações textuais. A composição das informações relevantes depende do tipo de modelagem: pode ser simulada tanto a reescrita integral do sumário, como normalmente é a tarefa manual (abstratos), quanto a simples justaposição de segmentos extraídos literalmente do texto original (extratos).
- b) Avaliação dos sumários para averiguar os resultados produzidos automaticamente:

- Conforme será visto mais adiante, a avaliação de sumários poderá ser realizada de forma intrínseca ou extrínseca.

Nas próximas Seções essas duas áreas de P&D de um sistema automático de sumarização serão detalhadas.

4.5.1.1 Modelagem de Procedimentos Automáticos para Escolha e Estruturação dos sumários

A área de PLN caracteriza duas abordagens distintas de sumarização automática:

- a) Abordagem superficial (ou empírica) – adota processamento prioritariamente baseado em reconhecimento de padrões de informações ou distribuições numéricas. São usadas técnicas empíricas e/ou estatísticas para a extração dos segmentos textuais relevantes. É capaz de produzir extratos textuais;
- b) Abordagem profunda (ou fundamental) – exige um alto grau de representação simbólica do conhecimento lingüístico e textual, e raciocínio, para estruturação e reescrita do sumário. Produzem abstratos.

A abordagem fundamental é de grande complexidade, uma vez que são envolvidas técnicas sofisticadas de PLN, como processos de interpretação e representação de textos. Tem um alto custo de P&D. Por outro lado, atualmente há um grande interesse pela abordagem empírica, que, diferentemente de incorporar modelos vinculados aos do comportamento humano na tarefa de sumarização, baseia-se em modelos matemáticos ou estatísticos para produzir resultados análogos (como aqueles vistos na seção 4.1.2). Tais modelos vêm sendo incrementados por recursos lingüísticos mais abrangentes, tornando os sumarizadores automáticos mais sofisticados e robustos. Por esse motivo, hoje em dia a adoção de métodos empíricos é muito promissora, principalmente com a necessidade de se processar grande volume de informação textual, disponível eletronicamente [RP2003].

O Modelo definido nesta Tese adota a abordagem superficial ou empírica de sumarização automática, sendo esta apresentada abaixo. Maiores informações sobre a abordagem fundamental podem ser encontradas em [Spark1993, BK1997, Marcu2000, MT1988].

A Abordagem Superficial: Sumarização Automática Extrativa

Segundo Espina e Rino [ER2002], métodos extrativos de SA consistem nas seguintes etapas:

- a) Identificar unidades de conteúdo que possam indicar a relevância dos segmentos textuais para a composição do sumário (termos representativos);
- b) Extrair do texto os segmentos mínimos que incluam tais unidades;
- c) Justapor cada um dos segmentos, resultando no sumário final.

As informações devem ser identificadas e delimitadas em segmentos textuais, reconhecendo-se tanto seu grau de relevância quanto seu inter-relacionamento. Essas características darão base às escolhas automáticas.

Em geral, os segmentos textuais relevantes são indicados por palavras ou elementos-chave mais complexos, cuja identificação é feita com base na análise da distribuição de frequência das palavras do texto fonte, considerando-se somente aquelas que possuem significados próprios, ou seja, as palavras de classe aberta (adjetivos, substantivos, advérbios e verbos). A etapa (a) acima é muito importante, pois dela depende, em grande parte, a qualidade dos extratos finais. A etapa (c) pode exigir, por exemplo, modificações da frase, tanto para adequação do português, bem como para torná-lo coeso.

A *coesão* pode ser definida como sendo a forma de certas palavras ou características gramaticais de uma sentença se conectar aos seus predecessores (e sucessores) em um texto. A coesão ocorre onde a interpretação de alguns elementos no discurso é dependente de outros. Por exemplo, para se entender uma referência “ele” em um texto, é preciso que se dê uma olhada em texto anterior que explicita a que “ele” se refere. Através da relação de coesão, duas cláusulas de texto são unidas [Barzilay1997].

São vários os métodos associados à abordagem superficial de SA, tendo suas pesquisas iniciadas no final da década de 50, explorando técnicas estatísticas simples para extração de sentenças significantes de textos. Até a década de 80, os métodos exploravam um subconjunto das seguintes técnicas [Luhn1958, Baxendale1958, Edmund1969, PZ1975, Paice1981, BJ1988]:

- Termos de maior frequência representam os conceitos mais importantes de um texto, correspondendo às chamadas *palavras-chave*;
- São consideradas palavras-chave apenas aquelas que carregam significados (classe aberta), como os substantivos, verbos, advérbios e adjetivos, exigindo então a remoção de *stopwords*: pronomes, preposições, artigos, conjunções e palavras irrelevantes no contexto;
- Cálculo da frequência das palavras-chave do texto para medir o grau de significância das diversas frases do texto. Sentenças mais significativas teriam pesos maiores, sendo escolhidas para compor o sumário, justapondo-as na ordem em que aparecem no texto;
- Palavras componentes de título e cabeçalho teriam maior probabilidade de serem representativas do tópico principal do texto, sendo a elas atribuído um peso maior;
- A posição da sentença no texto poderia associar um maior ou menor grau de importância. Por exemplo, frases no início e final do texto e frases logo abaixo de um título de seção poderiam receber maior peso. A primeira e última sentença de cada parágrafo poderia garantir conectividade com a anterior ou próxima sentença do texto (coesão textual);
- Uso de dicionário previamente construído de forma manual, contendo palavras consideradas relevantes no domínio do texto (palavras sinalizadoras), ou até mesmo frases (frase auto-indicativa). Por exemplo, em textos científicos poderiam ser consideradas importantes as palavras *conclusão* e *resultados*, e frases como *O objetivo deste artigo é*.

A partir da década de 90, impulsionado pelo surgimento de grandes *corpora* de textos *on-line* e tecnologias para sua manipulação eletrônica, nasce a área de mineração de texto (KDT) trazendo técnicas mais robustas de SA, fundamentadas em aprendizado e treinamento automáticos, além de outras técnicas como: aquisição de conhecimento lexical, revisão gramatical e tradução automática. Os avanços concentraram-se mais especificamente na análise textual, aparecendo os etiquetadores morfossintáticos e radicalizadores de palavras (*stemmer*). Os procedimentos estatísticos continuaram os mesmos, porém agora baseados em grandes *corpora* de domínios e gêneros variados, dando origem à metodologia baseada em *corpus* [MPER2001]. Os sistemas de SA dessa nova fase de evolução apresentaram, adicionalmente às técnicas já definidas anteriormente, as seguintes características [Marcu1997, Marcu2000, SSMB1997, KPC1995, TM1999, MH1991, LSKF2000]:

- Introdução de processo de *stemmer*;
- Palavras-chave anotadas com suas categorias morfossintáticas: sentenças com determinadas construções sintáticas e/ou morfológicas poderiam receber maior peso;
- Geração automática de critérios de ponderação e conjunto de parâmetros que caracterizariam o conteúdo textual, a partir de *corpora* de treinamento representativos de domínios específicos de textos, visando treinamento automático do sistema de SA. Após o treinamento, textos poderiam ser sumarizados a partir da seleção de sentenças de maiores pesos, baseados no conjunto de parâmetros e critérios de ponderação já disponibilizados;
- Parâmetros de características e critérios de ponderação poderiam compreender tanto características estatísticas como também morfossintáticas, utilizando-se então processos automáticos de etiquetagem para identificar grupos nominais (substantivos, adjetivos, preposições). Questões referentes à localização das sentenças e palavras/frases sinalizadoras também poderiam ser exploradas. Foram introduzidos também algoritmos de segmentação textual para identificar as cadeias léxicas mais fortes (coesão lexical);
- Introdução da medida de frequência inversa *tf-idf* (já vista anteriormente) para atribuição de pesos às palavras-chave. Esta medida traduz: palavras que ocorrem mais freqüentemente no documento do que no conjunto de treinamento recebem um peso maior, visto que estas diferenciarão este documento dos demais;
- Introdução da medida de frequência inversa de sentença *tf-isf* (*Term Frequency-Inverse Sentence Frequency*) para atribuição de pesos às sentenças. Este peso é similar à *tf-idf*, onde a noção de documento é substituída pela noção de sentença.

A simplicidade das técnicas envolvidas na abordagem superficial (mesmo complementado com técnicas de PLN) traz dificuldades na garantia da coesão e coerência dos sumários gerados [Sparck1993]. Entretanto, dependendo da aplicação a que se destinam tais sumários, pode ser interessante o uso dessas técnicas, como será mostrado nos resultados obtidos nesta Tese.

4.5.1.2 Avaliação dos Sumários Automáticos

As seções anteriores mostraram que, independente da abordagem adotada para a modelagem dos procedimentos automáticos de sumarização (superficial ou profunda), os sumários podem ser de diversos tipos e aplicabilidades, sendo, portanto, necessário se averiguar a adequação destes ao que se destinam.

Segundo Mani [Mani2001], não há consenso sobre as melhores métricas e métodos de avaliação de sistemas de SA, porém são vários os desafios a serem enfrentados. Por exemplo:

- Considerando as diversas perspectivas de uso de sumários e diferentes perfis de usuários (gostos, estilos, grau de conhecimento do assunto a ser sumarizado etc.), existe uma grande dificuldade na identificação do que seria um sumário adequado ou correto. Mesmo considerando usuários de perfis e perspectivas de uso semelhantes, um sumário pode ser considerado adequado por uns e inadequados por outros;
- Uma vez que a avaliação dos sumários é geralmente executada por humanos, este procedimento costuma ser caro, sendo preferível o uso de um modelo automático. Entretanto, dificilmente modelos automáticos substituem plenamente o julgamento humano. Além do mais, há escassez de especialistas em técnicas de sumarização, o processo é lento e complexo, e há um alto grau de subjetividade em julgamentos humanos;
- Como a sumarização envolve compressão, é importante avaliar os resultados sob diferentes taxas, o que aumenta a escala e complexidade do processo de avaliação. Além do mais, apesar do nível de informatividade do sumário ser diretamente proporcional ao seu tamanho, ele também é dependente do perfil do usuário, aplicabilidade, tempo disponível para lê-lo etc.

Segundo Sparck e Galliers [SG1996], métodos para avaliação de sumários podem ser classificados em duas grandes categorias: *intrínseca* e *extrínseca*. A avaliação intrínseca verifica o desempenho do sistema de sumarização, analisando principalmente a coerência e informatividade do sumário. A avaliação extrínseca testa se os sumários estão adequados a outras tarefas específicas, como categorização de documentos e recuperação de informação. Esses dois métodos foram utilizados na avaliação dos sumários produzidos pelo Modelo aqui proposto. Abaixo suas descrições.

A Avaliação Intrínseca

A coerência de um sumário é traduzida pela clareza de sua leitura, sendo, portanto, necessária a avaliação humana baseada em critérios específicos. Tais critérios podem envolver: presença de referências anafóricas não resolvidas, não preservação da integridade de estruturas como listas e tabelas, falta de coesão entre as sentenças do sumário, presença de tautologias (vício de linguagem que consiste em repetir o mesmo pensamento com palavras diferentes), ortografia, gramática, estilo etc. [Mani2001, SL2000, RP2003].

A informatividade de um sumário expressa o quanto do conteúdo original o sumário contém. A principal questão aqui é determinar a taxa de compressão ideal para não se correr o risco de excluir informação importante, bem como para não se incluir informação irrelevante. Entretanto, como já dito anteriormente, o julgamento da importância da informação é subjetivo, pois depende de características como perspectivas de uso, perfis de usuários e de juízes de avaliação.

A informatividade de um sumário pode ser avaliada por juízes humanos, comparando-se o conteúdo fonte com o sumário gerado, bem como por mecanismos automatizados, caso o texto fonte possua um sumário de referência. Neste último caso, quando o sumário de referência é do tipo extrativo, pode-se avaliar o resultado calculando-se sua precisão e cobertura (revocação). A precisão indica quantas sentenças do sumário de referência o sumário automático possui, em relação a todas as sentenças que ele contém. A cobertura indica quantas sentenças do sumário de referência o sumário automático possui, em relação a todas as sentenças que ele deveria possuir.

Existem algumas outras formas de avaliação automática e manual da informatividade de sumários, que envolvem, por exemplo: *medida de utilidade* - juízes humanos atribuem notas para todas as sentenças de um texto fonte [RJB2000]; *medida de conteúdo* - a parcela do conteúdo do sumário de referência que é transmitida pelo sumário automático, não levando em consideração valores numéricos (quantidade de sentenças ou tamanho dos sumários, por exemplo) [SM1983].

A Avaliação Extrínseca

A avaliação extrínseca de sumários automáticos verifica se estes estariam atendendo a perspectivas específicas de outras aplicações, como por exemplo: categorização de documentos, recuperação de informação e perguntas e respostas. Se o sumário envolve instruções de algum tipo, é possível também medir sua eficiência na execução dessas instruções. Este tipo de avaliação normalmente necessita de juízes humanos, sendo, portanto, custosa. Só é viável para documentos relativamente curtos, para facilitar o trabalho dos juízes.

Na categorização de documentos, o objetivo é verificar se os sumários automáticos apresentam informação suficiente para a correta classificação dos textos fonte. É avaliado tanto o tempo necessário para a realização da tarefa, como a taxa de acerto. Idealmente espera-se que o tempo para classificação a partir dos sumários seja significativamente menor que a classificação utilizando-se do texto fonte. Para a taxa de acerto, espera-se que esta seja tão alta quanto se a tarefa fosse realizada lendo-se os textos fonte.

Para a recuperação de informação, considerando-se uma base de dados com documentos que possuem sumários gerados automaticamente, o objetivo é verificar se tais sumários são suficientes para atender às buscas de usuários. Semelhante à categorização, é avaliado tanto o tempo necessário para a realização da tarefa, como a taxa de acerto.

Com relação a perguntas e respostas, o objetivo é verificar se juízes humanos conseguem responder a questões previamente elaboradas sobre o conteúdo do documento, com base unicamente no conteúdo do sumário automático. Perguntas que os juízes saberiam responder sem a leitura do documento (senso comum), ou que não conseguiriam responder mesmo com a leitura do documento (resposta não muito clara), são excluídas.

SIGAMeta - Um Modelo de Geração Automática de Metadados

Conforme visto no Capítulo 1, o Modelo SIGAMeta proposto nesta Tese converte automaticamente documentos textuais na língua portuguesa, no formato não estruturado (txt), em páginas *Web* anotadas semanticamente, utilizando padrões e tecnologias recomendados pela Web Semântica (RDF/XML e DC).

Este Capítulo define a arquitetura geral do Modelo SIGAMeta e seu funcionamento, interfaces gráficas, o conjunto de elementos de metadados, métodos e algoritmos para extração dos metadados, e métricas e métodos de avaliação. Os *corpora* de documentos textuais também são descritos.

Apesar de algumas informações abaixo já terem sido apresentadas no Capítulo 1, optou-se em reapresentá-las para efeito de praticidade na explicação deste Capítulo.

5.1 *Corpora* de Documentos Textuais a serem Processados pelo Modelo SIGAMeta

Os documentos textuais estudados nesta Tese foram recolhidos de forma aleatória em *sites* especializados nas categorias estudadas. Abaixo se encontra a descrição desses *corpora*.

5.1.1 *Corpora* de Treinamento em Categorias Específicas

Os *corpora de treinamento* foram compostos por 208 textos, com número aproximado de palavras variando entre 900 e 5.121. Estes *corpora* foram formados por quatro *corpus* de documentos textuais na língua portuguesa (formato txt), pertencentes às categorias Ciências Políticas (61 textos), Economia (31 textos), Estudos Populacionais (35 textos) e Antropologia Cultural (81 textos).

Estes documentos foram recolhidos respectivamente das seguintes fontes: Centro de Pesquisa e Documentação de História Contemporânea do Brasil – CPDOC, da Fundação Getúlio Vargas (<http://www.cpdoc.fgv.br/comum/htm/index.htm>); Biblioteca Virtual de Economia do Ministério da Ciência e Tecnologia (<http://www.prossiga.br/nuca-ie-ufrj/economia/>); Associação Brasileira de Estudos Populacionais – ABEP (<http://www.abep.org.br/usuario/GerenciaNavegacao.php>); Folclore Brasileiro – Lendas, Crenças, Danças, Brincadeiras Infantis e Simpatias (<http://www.ifolclore.com.br/index.htm>).

Os documentos de Ciências Políticas reuniram assuntos sobre a história da política, constituição brasileira, reforma judiciária, democracia, tribunal de contas, eleições governamentais, partidos políticos, governos de João Goulart, Juscelino Kubitschek, Getúlio Vargas e outros assuntos.

O *corpus* de Economia tratou de assuntos como custos e modos de produção, atividades econômicas de diversos municípios, crises econômicas e financeiras, economia brasileira, abertura comercial, competitividade, internacionalização do capital, inflação, comércio eletrônico, segurança financeira, investimentos etc.

Em Estudos Populacionais, os documentos se referiram principalmente a assuntos de migrações, saúde, mortalidade infantil e de idosos, maternidade e aleitamento, trabalho feminino, urbanização, juventude, raça, renda familiar, perfil socioeconômico etc.

O *corpus* de Antropologia Cultural reuniu documentos sobre o folclore brasileiro, com textos sobre lendas, crenças religiosas, simbolismos, danças e festas típicas, culinária, brincadeiras infantis, artesanato etc.

5.1.2 *Corpora* para Geração dos Metadados

Os *corpora para geração dos metadados* foram compostos por 119 textos com número aproximado de palavras variando entre 600 e 4.641. Estes *corpora*, igualmente aos *corpora* de treinamento, foram formados por quatro *corpus* de documentos textuais na língua portuguesa (formato txt), pertencentes às categorias Ciências Políticas (27 textos), Economia (21 textos), Estudos Populacionais (20 textos) e Antropologia Cultural (26 textos). Foram também considerados documentos pertencentes a outras categorias, objetivando verificar o

comportamento do Modelo diante de categorias não previstas (25 textos). Todos esses documentos foram recolhidos do *site* da Fundação Joaquim Nabuco - FJN (www.fundaj.gov.br), nas opções *Textos on-line* e *Pesquisa Escolar*.

Os *corpora* de Ciências Políticas da FJN reuniram documentos sobre políticos como: Miguel Arraes, João Pessoa, Epitácio Pessoa, Getúlio Vargas, Lula, Ronaldo Lessa, Jarbas Vasconcelos, Luiz Alberto Moniz e Agamenon Magalhães. Outros assuntos foram: partidos políticos, reforma agrária, eleições estaduais e democracia.

O *corpus* de Economia tratou de assuntos como: política de desenvolvimento regional, finanças sociais, trajetórias econômicas de municípios do norte e nordeste, trabalho e renda, serviços básicos, desemprego e exclusão social, dados sócio-econômicos etc.

Em Estudos Populacionais, os documentos se referiram principalmente à situação da mulher nordestina, tratando de assuntos como: saúde, mortalidade, emprego, fecundidade, urbanização, educação etc.

O *corpus* de Antropologia Cultural reuniu documentos sobre o folclore do norte e nordeste, com textos sobre lendas, religião, danças e festas típicas, superstições, brincadeiras infantis, artesanato, literatura de cordel, provérbios etc.

O *corpus* de outras categorias reuniu textos contendo bibliografias de personalidades do norte e nordeste do Brasil, pontos históricos do Recife e letras de músicas.

5.2 Elementos de Metadados Gerados pelo Modelo SIGAMeta

Os elementos de metadados gerados pelo Modelo seguiram o padrão *Dublin Core* [DublinCore, SVM2000], já definido e detalhado no Capítulo 3.

Como já mencionado no Capítulo 1, o Modelo aqui apresentado propõe a geração automática de nove elementos de metadados do DC (Tabela 5.1), e a geração semi-automática dos seis elementos restantes (Tabela 5.2).

Tabela 5.1 – Elementos de Metadados DC Gerados pelo Modelo de forma Automática

<i>Elemento</i>	<i>Conteúdo gerado</i>	<i>Técnica de geração</i>	<i>Tipo/função</i>
Subject	Palavras-chave e categoria do documento	Técnicas estatísticas, KDT e PLN	Descritivo
Description	Extrato	Técnicas estatísticas, KDT e PLN	Descritivo
Publisher	FJN	Parâmetro fornecido pelo usuário	Administrativo
Date	(AAAA-MM-DD) - Data de criação do documento digital	Fornecida pelo Sistema Operacional	Técnico
ObjectType	text	Parâmetro fornecido pelo usuário	Técnico
Format	text/xml	Parâmetro fornecido pelo usuário	Técnico
Identifier	Categoria+seqüencial	Parâmetro fornecido pelo usuário	Descritivo
Language	pt-br	Parâmetro fornecido pelo usuário	Descritivo
Rights	FJN	Parâmetro fornecido pelo usuário	Administrativo

O conteúdo gerado para os elementos de metadados acima estão de acordo com aqueles recomendados pelo DC (ver Tabela 3.3).

Tabela 5.2 – Elementos de Metadados DC Gerados pelo Modelo de forma Semi-automática

<i>Elemento</i>	<i>Conteúdo gerado</i>	<i>Técnica de geração</i>	<i>Tipo/função</i>
Title	“”	Semi-automática	Descritivo
Author	“”	Semi-automática	Descritivo
Contributor	“”	Semi-automática	Descritivo
Relation	“”	Semi-automática	Descritivo
Source	“”	Semi-automática	Descritivo
Coverage	“”	Semi-automática	Descritivo

A escolha dos elementos do DC para geração automática foi baseada em três fatores: 1) importância na retenção de contexto (semântica); 2) possibilidade de extração do elemento a partir do conteúdo do documento; 3) possibilidade de extração do elemento a partir de parâmetros fornecidos pelo usuário ou pelo sistema operacional.

Os elementos de metadados DC *Description* e *Subject* são os principais responsáveis em apreender a semântica de documentos textuais. Eles foram obtidos através da análise do conteúdo dos documentos, utilizando-se técnicas estatísticas, de mineração de textos e de processamento de linguagem natural. Os elementos restantes da Tabela 5.1, apesar de ausentes no conteúdo do documento, foram fornecidos pelo sistema operacional ou a partir

de parâmetros do usuário, sendo então replicados automaticamente para todos os documentos processados.

Para os elementos de metadados *Title*, *Author*, *Contributor*, *Relation*, *Source* e *Coverage*, o Modelo propôs a geração semi-automática. Os elementos *Title* e *Author*, apesar de presentes no conteúdo de documentos, não oferecem pistas seguras para a sua correta identificação automática, quando se tratando de documentos não estruturados. O restante desses elementos nem sempre está presente nos documentos e, por ser informação muito variada, não há como o usuário especificar parâmetro único que o represente em todos os documentos. Assim, para esses seis elementos de metadados do DC, o Modelo forneceu a sua codificação em RDF/XML, com conteúdo vazio, para serem posteriormente preenchidos de forma manual. Este procedimento semi-automático proporciona ao administrador de um ambiente uma maior flexibilidade na complementação de documentos digitais.

5.3 Funcionamento Geral do Modelo SIGAMeta e Interfaces Gráficas

O Modelo foi concebido para ser executado em duas fases: Fase de treinamento e Fase de geração dos metadados. Em ambas as fases, parâmetros de usuário foram coletados.

A Fase de treinamento teve o objetivo de extrair, a partir de cada *corpus* dos *corpora* de treinamento, os elementos que representaram cada categoria (vetor composto de termos, estatísticas e valores numéricos). Uma vez gerados os vetores das categorias, esta fase não mais necessitou ser repetida.

A Fase de geração dos metadados executou os algoritmos que extraíram, para cada documento, cada um dos nove elementos de metadados do DC gerados automaticamente, a partir dos *corpora* para geração dos metadados e dos parâmetros de usuários. Os vetores de categorias gerados na primeira fase foram utilizados unicamente para a geração do metadado *Subject* (Categorização por Diferença de Similaridade).

Uma vez extraídos os metadados, o Modelo converteu cada documento para o formato RDF/XML, complementando-os com os metadados a serem gerados de forma semi-automática. Para a execução do Modelo, uma interface gráfica de usuário foi desenvolvida tanto para unificar todo o funcionamento do sistema, como também para receber parâmetros de opções de usuários, indicação da localização dos arquivos a serem processados, manutenção das bases de dados geradas e bases de dados de parâmetros etc. Três módulos

principais foram definidos: Módulo de Configuração, Módulo de Extração de Padrões de Categorias e Módulo de Geração de Metadados. Abaixo se encontram as imagens das telas e suas descrições.

5.3.1 Módulo de Configuração

O Módulo de Configuração ofereceu meios para especificação e consulta de parâmetros globais do sistema (Figura 5.1).

Figura 5.1 – Interface Gráfica do Módulo de Configuração do Modelo

Descrição dos Parâmetros:

- Diretório de dados - diretório onde se encontram os arquivos de dados utilizados internamente pelo sistema;
- Arquivo de categorias - banco de dados das categorias em formato XML;
- Arquivo de *stopwords* - banco de dados de *stopwords* em formato texto;
- Extensão dos arquivos - extensão dos arquivos processados pelo sistema;
- Diferença máxima – similaridade - diferença máxima entre a similaridade do documento processado e a menor similaridade da categoria;
- Frequência mínima - menor frequência permitida para as palavras-chave;

- Peso mínimo - menor peso permitido para as palavras-chave incluídas nos vetores de categorias;
- Complexidade do termo composto - número máximo de termos das palavras-chave compostas;
- Diretório de saída dos metadados - diretório onde são armazenados os arquivos de metadados em formato RDF/XML;
- Diretório de saída dos resumos - diretório onde são armazenados os arquivos de resumos (extratos) em formato texto;
- Metadados – *Rights* - elemento DC *Rights* inserido nos arquivos de metadados (valor *default* FJN);
- Metadados – *Format* - Elemento DC *Format* inserido nos arquivos de metadados (valor *default* text/xml);
- Metadados – *Language* - Elemento DC *Language* inserido nos arquivos de metadados (valor *default* pt-br);
- Metadados – *ObjectType* - Elemento DC *ObjectType* inserido nos arquivos de metadados (valor *default* text);
- Metadados – *Publisher* - Elemento DC *Publisher* inserido nos arquivos de metadados (valor *default* FJN).

O parâmetro *Diferença máxima de similaridade* – DMS, como será explicado melhor mais adiante, foi utilizado apenas para a geração do metadado *Subject*, e teve o objetivo de especificar a diferença máxima aceita entre a menor similaridade da categoria e a similaridade do documento a ser categorizado (quando este obteve similaridade abaixo da menor similaridade da categoria).

5.3.2 Módulo de Extração de Padrões de Categorias

Este módulo teve o objetivo de executar a Fase de Treinamento do Modelo, extraíndo, a partir de cada *corpus* (dos *corpora* de treinamento), os elementos que representaram cada categoria. Sua interface gráfica encontra-se abaixo:

Figura 5.2 – Interface Gráfica do Módulo de Extração de Padrões de Categorias

Parâmetros utilizados:

- Diretório do *corpus* de treinamento - diretório em que se encontram os arquivos que serão processados para a geração da categoria;
- Palavras - número de palavras por documento a ser considerado no processamento;
- Identificador – faz parte do elemento de metadado DC *Identifier*. Conjunto de caracteres, geralmente a sigla da categoria, que será concatenada com um número seqüencial para identificar os documentos gerados sob a categoria correspondente, durante a criação dos metadados;
- Nome da categoria - nome atribuído à categoria processada;
- Processar *stemming* - opção que identifica o uso (ou não) do processador de *stemmer* (é possível escolher entre os algoritmos RSLP e Porter).

5.3.3 Módulo de Geração de Metadados

Este módulo permitiu a geração dos metadados em formato RDF/XML. A Figura 5.3 exibe sua interface gráfica.

Figura 5.3 – Interface Gráfica do Módulo de Geração de Metadados

Parâmetros utilizados:

- Diretório do *corpus* para geração de metadados - diretório em que se encontram os arquivos que serão processados para extração dos metadados;
- Palavras - número de palavras por documento a ser considerado no processamento;
- Compressão – taxa de compressão do documento para geração do sumário extrativo (metadado *Description*);
- Processar *stemming* - opção que identifica o uso do processador de *stemmer* (é possível escolher entre os algoritmos RSLP e Porter).

5.4 Arquitetura Geral e Detalhada do Modelo SIGAMeta

A Figura 5.4 apresenta o diagrama geral do Modelo. Este foi desenvolvido utilizando-se a linguagem de programação Java, produzido pela *Sun Microsystem Inc.* Como já dito na seção anterior, o Modelo foi concebido para ser executado em duas fases: Fase de treinamento e Fase de geração dos metadados.

Figura 5.4 – Diagrama Geral do Modelo

O Anexo A apresenta o Diagrama de Caso de Uso e o Diagrama de Sequência do Modelo, em linguagem UML (*Unified Modelling Language*).

5.4.1 Detalhamento da Fase de Treinamento

Como pode ser visto na Figura 5.4, a Fase de treinamento foi composta pelos elementos: *Corpora de treinamento em categorias específicas*, *Extração de padrões das categorias* e *Vetores de categorias*. Abaixo a descrição de cada um deles.

a) *Corpora* de treinamento em categorias específicas

Os *corpora de treinamento* foram compostos por documentos das categorias de Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural, conforme definidos na Seção 5.1.1. Antes de serem processados, esses documentos foram submetidos a um tratamento prévio manual, retirando-se dados sobre autores, referências bibliográficas, figuras, tabelas e gráficos. Através da interface gráfica do Módulo de Extração de Padrões de Categorias (Figura 5.2), pôde-se estabelecer a quantidade de palavras a serem processadas por documento e o diretório de cada *corpus* de treinamento.

b) Extração de padrões das categorias

O elemento *Extração de padrões das categorias* foi composto por dois componentes: pré-processamento e geração de características (Figura 5.5).

Figura 5.5 – Ampliação do Módulo de Extração de Padrões das Categorias

Pré-processamento:

O *pré-processamento* envolveu atividades de *case folding*, *stopwords* e *stemmer*.

- *Case Folding* é o processo de padronização das letras do texto (maiúsculas ou minúsculas), trazendo a vantagem de acelerar comparações no processo de indexação. Neste passo foram também retirados caracteres especiais, pontuações e números;

- A lista de *stopwords* foi composta por 2.679 termos criados manualmente, compreendendo: palavras sem conteúdo semântico no contexto, palavras auxiliares ou conectivas, preposições, pronomes, artigos, alguns verbos (como ser e estar) e outros (ver no Anexo B, Seção B.1, algumas dessas palavras);
- Para o processo de *stemmer*, o Modelo ofereceu duas opções de algoritmos: Porter [Porter1980] e RLSP [Orengo2004, OH2001]. A especificação da escolha do algoritmo pôde ser feita através da interface gráfica do usuário.

Geração de características:

O componente *Geração de características* analisou cada documento que compôs a categoria, selecionando termos simples (por exemplo, “computação”), duplos (por exemplo, “Ciência computação”) e triplos (por exemplo, “linguagem programação Java”) que foram considerados representativos da categoria. O parâmetro *Complexidade do termo composto* do Módulo de Configuração (Figura 5.1) foi o responsável em especificar o número máximo de termos consecutivos (nesses experimentos foi especificado valor três).

A atribuição de significância aos termos (peso) foi realizada através da fórmula F07 abaixo [CMNS1996], que é uma variação do cálculo da frequência relativa normalizada, referenciada como F06 no Capítulo 4. Esta alteração é uma tentativa de valorizar termos formados por mais de uma palavra, pois estes geralmente carregam uma semântica mais precisa que os termos simples. Palavras cuja frequência absoluta (tf_{ik}) foi menor que três (3) foram excluídas. Este valor foi adotado pela observação da quantidade e qualidade das palavras-chave geradas com valores inferiores e superiores, em todas as categorias. O Módulo de Configuração, através do parâmetro *Frequência mínima*, fez o controle dessa variável.

$$w_{ik} = \frac{tf_{ik} \log(N / n_k) * P_{tk}}{\sqrt{\sum_{k=1}^t (tf_{ik})^2 [\log((N/n_k) * P_{tk})]^2}} \quad (F07)$$

Nesta fórmula:

- W_{ik} representa o peso do termo k no documento i ;
- tf_{ik} representa a frequência do termo k no documento i ;
- N é a quantidade de documentos no *corpus* C ;
- n_k é a quantidade de documentos que possuem o termo T_k ;
- t é o número de termos no documento;
- P_{tk} é o número de palavras que compõe o termo T_k .

Observar que quando P_{tk} é igual a 1 (um), a fórmula acima é a mesma F06, apresentada no Capítulo 4.

Para seleção dos termos relevantes de cada documento, adotou-se a técnica de truncagem, selecionando-se termos cujos pesos foram maiores ou iguais a 0,002. Este valor foi adotado pela observação da qualidade das palavras-chave geradas com pesos inferiores a este valor em todas as categorias. Observou-se também que o número de documentos que continham essas palavras-chave (n_k) era igual a 1. O parâmetro *Peso mínimo* do Módulo de configuração controlou essa variável.

c) Vetores de categorias

Uma vez calculados os pesos e selecionados os termos representativos de cada documento da categoria, gerou-se o *vetor de categoria*, composto pelos seguintes elementos:

- Nome da categoria;
- Termos selecionados para representar a categoria (*stemmings*);
- Média dos pesos desses termos;
- Número de documentos que possui cada um dos termos (n_k);
- Tamanho do *corpus* (N);
- A menor similaridade encontrada entre um documento do *corpus* e o vetor de categoria gerado (denominado *Sim-min*).

O cálculo de similaridade foi realizado através da fórmula do co-seno (ou produto interno normalizado), apresentada no Capítulo 4 (F01). Todas essas informações foram necessárias para o processamento da categorização dos documentos na segunda fase do Modelo (Geração de metadados). A Fase de treinamento foi realizada por quatro vezes,

gerando os vetores de categorias de Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural.

5.4.2 Detalhamento da Fase de Geração de Metadados

A Fase de geração de metadados foi composta pelos elementos: *Corpora para geração de metadados*, *Vetores de categorias*, *Geração de metadados*, *Conversão RDF/XML* e *Documentos em RDF/XML*. Abaixo a descrição de cada um deles.

a) *Corpora* para geração de metadados

Os *corpora para geração dos metadados*, igualmente aos *Corpora* de treinamento, foram compostos por documentos das categorias de Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural, conforme definidos na Seção 5.1.2. Foram também considerados documentos pertencentes a outras categorias (Outras) para verificar o comportamento do Modelo diante de categorias não previstas.

b) Vetores de categorias

Este elemento foi formado pelo conjunto de vetores de categorias de Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural, gerados na Fase de treinamento do Modelo.

c) Geração de metadados

O elemento *Geração de metadados* foi formado por cinco componentes: pré-processamento, atribuição de pesos e seleção de termos, categorização, sumarização e geração de outros metadados (Figura 5.6).

Figura 5.6 – Ampliação do módulo de Geração de Metadados

Pré-processamento:

O componente *pré-processamento* teve os mesmos elementos e comportamento definidos para o seu correspondente na Fase de treinamento, porém agora operando com os *Corpora* para geração de metadados.

Atribuição de pesos e seleção de termos:

O componente *Atribuição de pesos e seleção de termos* analisou cada documento selecionando termos simples, duplos e triplos (consecutivos) considerados significantes para o documento (parâmetro *Complexidade do termo composto* do Módulo de Configuração igual a três). A atribuição de significância aos termos foi realizada através de sua frequência absoluta (tf_{ik}), apresentada no Capítulo 4. Palavras com frequência menor que três (3) foram excluídas (parâmetro *Frequência mínima* igual a três). Este cálculo de peso foi preliminar, sendo complementado de forma específica nos processos de categorização e sumarização.

Categorização:

Para a categorização dos documentos, o Modelo definiu a seguinte estratégia, atribuindo-lhe o nome de *Categorização por Diferença de Similaridade*:

- Partiu-se do princípio de que o documento a ser categorizado poderia pertencer a qualquer uma das quatro categorias previstas pelo Modelo (Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural), ou até mesmo a nenhuma delas;
- Para cada uma das categorias, calculou-se o peso de cada termo-chave do documento pela fórmula F07 (exibida na primeira fase); calculou-se a similaridade de cada documento com cada um dos quatro vetores de categoria ($Sim(doc, VC)$), utilizando-se a fórmula do co-seno apresentada no Capítulo 4 (F01); subtraiu-se a similaridade obtida ($Sim(doc, VC)$) da similaridade mínima ($Sim-min$) de cada vetor de categoria, gerando-se um valor X para cada categoria (X é a diferença de similaridade);
- Caso houvesse valores X maiores ou iguais a zero, considerou-se o mais alto. Se todos os resultados de X fossem negativos, assumiu-se o menor valor absoluto. Para este último caso, considerou-se também o parâmetro *Diferença máxima de similaridade* – DMS, especificado pelo usuário na interface gráfica (*Módulo de Configuração*). Documentos cujos resultados estivessem acima deste valor limite foram considerados de categoria desconhecida.

A idéia principal de se usar o parâmetro DMS é a tentativa de classificar corretamente documentos cujas similaridades fossem um pouco abaixo que a menor similaridade de suas respectivas categorias ($Sim-min$). Uma vez que os documentos que fizeram parte dos *corpora* de treinamento foram selecionados de forma aleatória, sem se conhecer a fundo seus

conteúdos (mas apenas a categoria a qual pertenciam), não se pôde garantir que estes continham documentos mínimos que gerassem a menor similaridade possível para a categoria.

Observa-se, entretanto, que este valor não pode ser muito alto, pois assim estar-se-ia fadado a classificar indevidamente os documentos. Dessa forma, nos experimentos apresentados aqui, a categorização foi avaliada para alguns valores desse parâmetro (0,0; 0,01 e 0,02), buscando-se achar um valor ideal específico para o *corpora* de geração de metadados adotado neste trabalho.

Sumarização:

Para a *Sumarização* dos documentos (sumarização extrativa) utilizou-se a junção de algumas estratégias já empregadas em alguns experimentos [RP2003, BLC2000, Marcu2000, SSMB1997, KPC1995, TM1999, LSKF2000].

A atribuição de pesos aos termos foi realizada pela normalização de sua frequência absoluta (tf_{ik}), calculada no componente *Atribuição de pesos e seleção de termos*. Esta normalização utilizou a fórmula F03, apresentada no Capítulo 4. Para cada uma das sentenças do documento foi atribuído um peso que correspondeu à média dos pesos dos termos que formaram a sentença. Para a seleção das sentenças componentes do sumário, utilizou-se o esquema abaixo:

- A sentença que gerou o maior peso foi considerada a mais importante do documento, sendo esta sempre escolhida para compor o sumário (*gist sentence*);
- Cada sentença selecionada para o sumário teve no mínimo peso igual à média dos pesos de todas as sentenças do documento;
- Cada sentença selecionada para o sumário teve pelo menos um termo-chave da *gist sentence* (visando obter uma maior coesão de sentenças);
- O número de sentenças selecionadas obedeceu a uma taxa de compressão definida na interface gráfica do Módulo de Geração de Metadados (parâmetro *Compressão*). Isso foi feito eliminando-se as sentenças de menor peso, dentre as sentenças selecionadas;

- No final do procedimento, caso o sumário ainda comportasse frases, visando uma maior coesão, foram incorporadas sentenças localizadas antes e depois das sentenças escolhidas, desde que dentro do mesmo parágrafo e respeitando-se a taxa de compressão. Neste caso, não foram respeitadas as condições dos itens segundo e terceiro acima;
- As sentenças foram justapostas seguindo-se a mesma ordem encontrada no documento.

Geração de outros metadados:

Os demais metadados foram gerados a partir de parâmetros fornecidos pelo usuário na interface gráfica Módulo de Configuração (ver Seção 5.3.1). Os elementos de metadados *Publisher*, *ObjectType*, *Format*, *Language* e *Rights* foram preenchidos com os valores especificados na Tabela 5.1. Os elementos *Title*, *Author*, *Contributor*, *Relation*, *Source* e *Coverage*, definidos para geração semi-automática, foram preenchidos com conteúdo vazio.

O metadado *Identifier* foi composto pelo conteúdo do parâmetro *Identificador* fornecido pelo usuário no Módulo de Extração de Padrões de Categorias, concatenado a um número sequencial gerado automaticamente pelo Modelo para cada uma das categorias. O metadado *Date* foi gerado a partir de informação do sistema operacional.

d) Conversão RDF/XML

A conversão dos documentos consistiu em gerar, para cada um dos documentos processados, o seu correspondente no formato RDF/XML, e metadados no padrão *Dublin Core*. Cada documento RDF/XML foi validado por *sites* disponibilizados pela Internet, invocados em suas primeiras linhas:

```
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:dc='http://purl.org/dc/elements/1.1/'
xmlns:dcterms='http://purl.org/dc/terms/'>
```

Para a formatação de exibição, um documento XSL (*eXtensible Stylesheet Language*) foi desenvolvido (ver Anexo B, seções B.6 e B.7).

e) Documentos em RDF/XML

Os documentos RDF/XML validados foram armazenados em uma base de dados para serem avaliados.

5.5 Métricas e Métodos de Avaliação do Modelo SIGAMeta

Uma vez que o objetivo desta Tese é “fornecer uma contribuição na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*”, a finalidade principal da avaliação do Modelo foi verificar (através das métricas de precisão e revocação), o comportamento dos documentos convertidos para o formato RDF/XML, quando estes foram submetidos a um processo de recuperação de informação. Os elementos de metadados *Description* e *Subject* foram exaustivamente avaliados, uma vez que estes são os principais responsáveis em apreender a semântica de documentos textuais. Além do mais, estes necessitaram de algoritmos e métodos complexos de extração de informação.

Apesar de não ser objetivo principal dessa Tese a avaliação de desempenho dos algoritmos de *stemmer* Porter e RSLP utilizados nesse experimento, estes foram avaliados em termos de influência na geração dos metadados *Description* e *Subject*. Vale a pena ressaltar que não se encontrou trabalhos que tenham realizado avaliações de algoritmos de *stemmer* no âmbito da categorização e da sumarização de textos. Também foi avaliada a qualidade dos *stemmings* gerados.

Uma vez que não se dispunha de um aplicativo de RI com as características necessárias para a avaliação de alguns resultados dessa Tese, desenvolveu-se um em linguagem de programação Java, cujas características encontram-se descritas abaixo (ver interface gráfica na Figura 5.7):

- Aceita as opções “todos os termos” ou “qualquer termo”, através de uma interface gráfica;
- Verifica a validade da expressão de busca no conteúdo dos metadados *subject* e *description*, para o caso de documentos em RDF/XML;
- Verifica a validade da expressão de busca no conteúdo completo dos documentos, no caso de documentos em formato txt;

- Apresenta ao usuário duas listas de referências aos documentos encontrados, uma para as pesquisas nos documentos RDF/XML, e outra para as pesquisas nos documentos *txt*.
- Para cada uma das listas, apresenta também um resumo contendo a quantidade de documentos investigados, quantidade de documentos selecionados e o tempo gasto para processar o *corpus* completo.

Figura 5.7 – Interface Gráfica do Aplicativo de Recuperação de Informação

Os próximos itens descrevem os métodos de avaliação e métricas adotadas para o experimento.

a) Avaliação dos Documentos RDF/XML

A avaliação dos documentos RDF/XML adotou a seguinte estratégia:

- Submissão dos *corpora* para geração dos metadados ao Modelo proposto, gerando documentos em RDF/XML com metadados DC;
- Submissão dos documentos RDF/XML e dos documentos em formato *txt* ao aplicativo de RI ;

- Avaliação dos resultados utilizando-se as métricas de *precisão* (número de itens relevantes recuperados dividido pelo número total de itens recuperados) e de *revocação* (número de itens relevantes recuperados dividido pelo número de itens relevantes na coleção). Além destas, foi também investigado o tempo de resposta de cada caso.

b) Avaliação do Metadado *Description*

O metadado *Description* foi obtido através de técnicas de sumarização extrativa de textos, sendo sua avaliação realizada pela verificação dos seguintes elementos: coerência, preservação da idéia central e capacidade de gerar categorias corretas [SG1996]. Todos esses elementos foram avaliados para diferentes taxas de compressão atribuídas aos textos-fonte (70%, 80% e 90%).

A “coerência” do sumário foi avaliada por humanos, baseada principalmente no critério de clareza de conteúdo e coesão entre as sentenças do sumário. A “preservação da idéia central” também foi avaliada por humanos, comparando-se o conteúdo do documento fonte com o sumário gerado.

Para o elemento “capacidade de gerar categorias corretas”, o objetivo foi verificar se os sumários possuíam informação suficiente para expressar a categoria do documento. Para tanto, um conjunto de sumários gerados antecipadamente pelo Modelo foi submetido ao módulo de categorização, e avaliada a taxa de acerto e o tempo de categorização.

c) Avaliação do Metadado *Subject*

O metadado *Subject* foi obtido através da técnica de categorização assistida, onde primeiramente o Modelo foi “ensinado” a identificar características específicas das categorias previstas (Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural). Após a geração dessas características, adotou-se a estratégia de categorização definida nesta Tese como *Categorização por Diferença de Similaridade*.

A sua avaliação se baseou na taxa de acerto da geração das categorias dos documentos. Documentos de categorias não “ensinadas” ao Modelo (Outras) também foram processados para verificar seu comportamento. Esse experimento foi realizado para diferentes quantidades de palavras processadas por texto e para valores diferentes de

Diferença Máxima de Similaridade, quando processados com os algoritmos de *stemmer* RSLP e Porter.

d) Avaliação de Desempenho dos Algoritmos de *Stemmer*

Na comparação de algoritmos de *stemmer*, tradicionalmente os experimentos medem a efetividade de desempenho baseada no percentual de redução de palavras a seu radical. Nesta Tese, além de aplicar este tipo tradicional de avaliação, os algoritmos Porter e RSLP também foram avaliados quanto à qualidade desses radicais e quanto ao seu desempenho perante resultados de categorização e sumarização dos textos.

Avaliação e Análise dos Resultados

Uma vez definido o Modelo SIGAMeta para Geração Automática de Metadados no capítulo anterior, este capítulo vem apresentar os resultados e análises dos testes desse Modelo, com base nas métricas e métodos já definidos.

Para os experimentos descritos neste Capítulo, alguns parâmetros do Módulo de Configuração permaneceram inalterados. Foram eles:

- Frequência mínima dos termos de cada documento igual a três (parâmetro *Frequência mínima*). Este valor foi adotado com base na qualidade e quantidade de palavras-chave geradas para valores superiores e inferiores a 3;
- Peso dos termos maior ou igual a 0,002 (parâmetro *Peso mínimo*), apenas para a geração dos vetores das categorias. Este valor foi adotado pela observação da qualidade das palavras quando possuíam valores inferiores. Observou-se também que para os quatro vetores de categoria, estas palavras só ocorriam em um único documento;
- Número máximo de palavras consecutivas igual a três (parâmetro *complexidade de termo composto*).

6.1 Resultados Produzidos pelo Módulo de Extração de Padrões de Categorias

Os experimentos tiveram início com a execução do *Módulo de Extração de Padrões de Categorias* (Figura 5.2), Fase de Treinamento, extraíndo-se, a partir dos *corpora* de treinamento, os elementos simples, duplos e triplos (consecutivos) que representaram cada categoria. Antes de serem processados, todos os documentos foram submetidos a um tratamento prévio, retirando-se dados sobre autores, referências bibliográficas, figuras, tabelas e gráficos. A Tabela 6.1 exhibe as características dos *corpora* de treinamento.

Tabela 6.1 – *Corpora* de Treinamento em Categorias Específicas

<i>Nome da Categoria</i>	<i>Sigla</i>	<i>Número de documentos</i>	<i>Tamanho em (KB)</i>	<i>Tamanho médio (KB) por documento</i>	<i>Número aproximado de palavras por documento</i>
Ciências Políticas	CPOLI	61	397	6,5	900
Economia	ECONO	31	992	32,0	5.000
Estudos Populacionais	ESTPOP	35	1.140	32,6	5.121
Antropologia Cultural	ANTROP	81	700	8,64	1.300
Total		208	3.229		

Como pode ser visto nessa Tabela, os *corpora* de treinamento totalizaram 208 textos, com número aproximado de palavras variando entre 900 (Ciências Políticas) e 5.121 (Estudos Populacionais). Com base nesses dados, para a geração de cada vetor de categoria, foi especificado na interface gráfica o valor de 5.100 palavras a serem processadas por documento. Este processo foi realizado tanto para o *stemmer* RSLP como para o Porter.

Como dito no capítulo anterior (Seção 5.4.1), o vetor de categoria foi composto pelos seguintes elementos:

- Nome da categoria;
- Termos selecionados para representar a categoria (*stemmings*);
- Média dos pesos desses termos (calculado pela fórmula F07 descrita no capítulo anterior);
- Número de documentos que possui cada um dos termos (n_k);
- Tamanho do *corpus* (N);
- A menor similaridade encontrada entre um documento do *corpus* e o vetor de categoria gerado (*Sim-min*), calculada pela fórmula F01, mostrada no Capítulo 4.

As tabelas 6.2, 6.3, 6.4, 6.5 e 6.6 abaixo mostram essas informações quando utilizando os algoritmos de *stemmer* RSLP e Porter.

Tabela 6.2 – Características dos Vetores de Categorias

<i>Nome da Categoria</i>	<i>Sigla</i>	<i>Número de documentos processados (N)</i>	<i>Número de termos selecionados</i>	<i>Menor similaridade (Sim-min)</i>
Ciências Políticas	CPOLI - RSLP	61	390	0,019601
	CPOLI - Porter		392	0,021683
Economia	ECONO – RSLP	31	1.008	0,056866
	ECONO - Porter		1.027	0,052349
Estudos Populacionais	ESTPOP – RSLP	35	824	0,130573
	ESTPOP - Porter		834	0,119261
Antropologia Cultural	ANTROP – RSLP	81	489	0,017749
	ANTROP - Porter		500	0,017312

O método de seleção dos termos das categorias foi explicado no capítulo anterior, Seção 5.4.1. A diferença no número de termos selecionados para Porter e RSLP ocorre tanto por causa da forma de geração dos *stemmings*, específico de cada algoritmo, como também por causa do valor do peso calculado para cada *stemming* (lembrar do ponto de corte adotado: peso $\geq 0,002$).

É interessante observar que de todas as categorias, a menor similaridade gerada foi para Antropologia Cultural, seguida de Ciências Políticas. A maior similaridade ficou para Estudos populacionais. Isto significa dizer (como será comprovado mais adiante) que documentos poderão alcançar mais facilmente uma similaridade com Antropologia Cultural e Ciências Políticas, do que com Estudos Populacionais.

Para entender melhor o cálculo de similaridade, é importante lembrar que ele não é trivial. Sua fórmula (F01, com pesos calculados pela F07) leva em consideração não apenas a frequência dos termos no documento a ser categorizado, mas também informações sobre o comportamento desses termos nos documentos do *corpus* da categoria que se está investigando. O número de palavras por documento também influenciam no sentido de que gerarão um número maior de palavras-chave. Observando-se na Tabela 6.1 o número aproximado de palavras por documento, vê-se que as categorias Antropologia Cultural e Ciências Políticas também contiveram os menores documentos.

Para efeito de ilustração, as tabelas 6.3, 6.4, 6.5 e 6.6 exibem os seis primeiros *stemmings* de maior peso para cada uma das categorias geradas, quando se utilizou o

stemmer RSLP. No Anexo B (Seção B.2) encontram-se listas mais completas de *stemmings* desses vetores para RSLP e para Porter.

Tabela 6.3 – Seis Primeiros *Stemmings* da Categoria CPOLI - RSLP

<i>Stemming</i>	<i>Algumas palavras</i>	<i>Número de documentos (n)</i>	<i>Peso médio</i>
polit	político, política, politização	45	0,22462040261336310
govern	governo, governar, governista, governante, governador	38	0,18846437461276475
brasil	Brasil, brasileiro, Brasília	25	0,10318175371796491
presid	presidente, presidencial, presidência, presidir, presidencialista	26	0,10167994098086017
est	estado, estados	23	0,08966530078194465
trabalh	trabalhar, trabalho, trabalhador, trabalhista	21	0,08750279572420754

Tabela 6.4 – Seis Primeiros *Stemmings* da Categoria ECONO - RSLP

<i>Stemming</i>	<i>Algumas palavras</i>	<i>Número de documentos (n)</i>	<i>Peso médio</i>
econom	economia, econômico, economicamente, economizar, economista	16	0,11561841935294147
capit	capital, capitalismo, capitalista, capitalização	16	0,08972888985974249
empres	empresa, empresário	14	0,08375685801320701
país	país, países	14	0,06943114793005972
analís	analista, análise, analisado	14	0,06760117036058592
desenvolv	desenvolve, desenvolver, desenvolvido, desenvolvimento	16	0,062960833774728

Tabela 6.5 – Seis Primeiros *Stemmings* da Categoria ESTPOP - RSLP

<i>Stemming</i>	<i>Algumas palavras</i>	<i>Número de documentos (n)</i>	<i>Peso médio</i>
mulh	mulher, mulheres	22	0,16994960013369317
trabalh	trabalhar, trabalho, trabalhador, trabalhista	25	0,16178526418252928
analís	analista, análise, analisado	22	0,07465776273336182
migr	migra, migrar, migração, migrante, migratório	11	0,07242701821959494
popul	popular, população, populoso	13	0,06911143741470743
famil	família, famílias	13	0,06790235104345187

Tabela 6.6 – Seis Primeiros *Stemmings* da Categoria ANTROP - RSLP

<i>Stemming</i>	<i>Algumas palavras</i>	<i>Número de documentos (n)</i>	<i>Peso médio</i>
Danç	dança, dançar, dançante, dançando	30	0,16412275856894537
Cult	cultura, cultural, culto, cultivo	21	0,06485894906681130
Fest	festa, festivo, festeiro	19	0,06356523888107604
Folclor	folclore, folclórico, folclorista	14	0,05893565613347529
Pov	povo, povos	17	0,05323351718432739
Cant	cantar, canto, cantador, cantoria	19	0,04939378090182284

Como pode ser notado nas tabelas acima (e no Anexo B), o método adotado pelo Modelo para a seleção de termos das categorias gerou *stemmings* que realmente podem ser considerados representativos. O peso desses *stemmings* é a média de seus pesos em cada um dos documentos da categoria (calculado pela fórmula F07 descrita no capítulo anterior). A fórmula F07 gera pesos normalizados com valores entre zero e um.

Para avaliação de desempenho dos algoritmos de *stemmer* Porter e RSLP, conforme explicado no capítulo anterior (Seção 5.5), o trabalho propôs três formas: a) observação do percentual de redução de palavras a seu radical; b) observação da qualidade dos radicais e c) influência nos resultados de categorização e sumarização dos textos.

Para a avaliação do percentual de redução de palavras a seu radical, foram observados os termos gerados pelo componente *pré-processamento*, do Módulo *Extração de Padrões das Categorias*, a partir dos *corpora* de treinamento. Foram contabilizados o número de palavras distintas antes do processo de *stemmer* e o número de *stemmings* gerados por estas palavras para Porter e para RSLP. Os resultados obtidos encontram-se na Tabela 6.7.

Tabela 6.7 – Percentual de Redução de Palavras a seu *Stemming* (Porter X RSLP)

	<i>CPOLI</i>	<i>ECONO</i>	<i>ESTPOP</i>	<i>ANTROP</i>	<i>Total</i>
<i>Número de Palavras</i>	6.301	7.714	6.755	11.692	32.462
<i>Stemmer RSLP</i>	3.647	4.022	3.563	6.228	17.460
<i>Taxa de redução</i>	42,12%	47,86%	47,25%	46,73%	46,22%
<i>Stemmer Porter</i>	4.024	4.525	3.995	7.051	19.595
<i>Taxa de redução</i>	36,13%	41,34%	40,85%	39,69%	39,64%

Como pode ser visto na Tabela 6.7, quantitativamente o desempenho geral do algoritmo RSLP foi superior ao do algoritmo Porter, com taxa média de redução respectivamente de 46,22% e 39,64%. Entretanto, é necessário observar a qualidade dos *stemmings* gerados. Abaixo se encontram algumas palavras que geraram *stemmings* diferentes. Uma lista mais completa desses *stemmings* encontra-se no Anexo B (Seção B.3).

Tabela 6.8 – Exemplos de *Stemmings* RSLP e Porter

<i>Palavra</i>	<i>Stemming RSLP</i>	<i>Stemmings Porter</i>
abolição	abol	aboliçã
abolir	abol	abol
aboliu	abol	abol
academia	academ	academ
acadêmica	academ	acadêm
academicamente	academ	academ
cidadão	ciudad	ciudadã
cidadãos	ciudad	ciudadã
cidade	ciudad	ciudad
cidades	ciudad	ciudad
concorrência	concorr	concorrent
concorrencial	concorr	concorrencial
concorrendo	concorr	concorr
concorrente	concorr	concorrent
concorrer	concorr	concorr
culta	cult	cult
cultivo	cult	cultiv
culto	cult	cult
cultura	cult	cultur
cultural	cult	cultural

Como pode ser visto na Tabela 6.8, o algoritmo Porter gera um número maior de radicais diferentes do que o RSLP. Esta diferença se dá principalmente por causa da forma como ele trata palavras com acentuação. Observa-se, entretanto, que este comportamento pode vir a gerar *stemmings* de melhor qualidade ou não. Por exemplo, na Tabela 6.8, o grupo de palavras {*cidadão*, *cidadãos*, *cidade* e *cidades*} gerou o radical *ciudad* para RSLP e os radicais {*ciudadã* e *ciudad*} para Porter. Neste caso, o algoritmo RSLP cometeu erro de *overstemming* (remoção de caracteres que na realidade não fazem parte do sufixo, e sim do radical), uma vez que o radical *ciudad* não poderia representar palavras de significados

diferentes (cidadão e cidade). Neste caso, o algoritmo Porter ofereceu resultado de melhor qualidade. Este mesmo tipo de erro foi cometido também para o grupo de palavras {*culta, cultivado, culto, cultura, cultural*}.

Observando agora o grupo de palavras {*concorrência, concorrencial, concorrendo, concorrente, concorrer*}, vê-se que o algoritmo RSLP gerou radical de melhor qualidade (*concorr*) do que o algoritmo Porter (*concorrent, concorrencial, concorr*), que cometeu erro de *understemming* (não remoção completa do sufixo). Isso tudo justifica o fato do algoritmo RSLP ter gerado uma menor quantidade de *stemmings* que o algoritmo Porter.

Entretanto, comparando-se os dois tipos de erros, pode-se afirmar que erros de *overstemming* podem ser bem mais prejudiciais que erros de *understemming*, quando se tratando de aplicações que buscam extrair o contexto de informações. Dessa forma, conclui-se que o algoritmo Porter obteve um melhor desempenho na geração de *stemmings* de melhor qualidade que o algoritmo RSLP.

Mais adiante será avaliada a influência desses *stemmings* nos resultados de categorização e sumarização de textos.

6.2 Resultados Produzidos pelo Módulo de Geração de Metadados

Uma vez terminada a Fase de Treinamento, deu-se início à segunda fase do Modelo, executando-se o *Módulo de Geração de Metadados* (Figura 5.3).

Esse Módulo, conforme especificado na Seção 5.4.2, extraiu para cada documento dos *corpora* para geração dos metadados, cada um dos nove elementos de metadados do DC gerados automaticamente. Os vetores de categorias gerados na primeira fase foram utilizados unicamente para a geração do metadado *Subject*.

Diferentemente dos *corpora* de treinamento, por questões de custo de manutenção, os documentos dos *corpora* para geração dos metadados não sofreram nenhuma alteração antes de serem processados. A Tabela 6.9 exibe as características desses *corpora*.

Tabela 6.9 – *Corpora* para Geração dos Metadados

<i>Nome da Categoria</i>	<i>Sigla</i>	<i>Número de documentos</i>	<i>Tamanho em (KB)</i>	<i>Tamanho médio (KB) por documento</i>	<i>Número aproximado de palavras por documento</i>
Ciências Políticas	FJN-CPOLI	27	683	25,3	3.730
Economia	FJN-ECONO	21	607	28,9	4.300
Estudos Populacionais	FJN-ESTPOP	20	664	33,2	4.641
Antropologia Cultural	FJN-ANTROP	26	192	7,38	1000
Outras	FJN-OUTRAS	25	112	4,48	600
Total		119	2.258		

Como pode ser visto na Tabela 6.9, os *corpora* para geração dos metadados totalizaram 119 textos com um número aproximado de palavras variando entre 600 (categoria Outras) e 4.641 (categoria Estudos Populacionais).

6.2.1 Avaliação do Metadado *Subject*

O primeiro metadado a ser analisado foi o *Subject*, obtido através da estratégia de categorização por diferença de similaridade. Para tanto, uma vez gerados os vetores das categorias Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural, o *Módulo de Geração de Metadados* foi executado para todos os 119 textos descritos na Tabela 6.9, utilizando-se os *stemmer* Porter e RSLP. A sua avaliação se baseou na taxa de acerto na geração das categorias dos documentos, o que foi feito variando-se a quantidade de palavras a serem processadas por documento, quando o parâmetro *diferença máxima de similaridade* – DMS, assumiu valores 0,0; 0,01 e 0,02.

Como dito no capítulo anterior, a idéia principal de se usar o parâmetro DMS é a tentativa de classificar corretamente documentos cujas similaridades fossem um pouco abaixo que a menor similaridade de suas respectivas categorias (*Sim-min*). Observa-se, entretanto, que este valor não pode ser muito grande, para que não se classifique indevidamente o documento.

Antes de se avaliar os resultados da categorização, é interessante conhecer algumas características das cinco categorias processadas, para que se entendam melhor os resultados obtidos. Independente do algoritmo de *stemmer* utilizado (Porter ou RSLP), os documentos

categorizados erradamente pelo Modelo seguiram o seguinte padrão, de acordo com o valor da *Diferença Máxima de Similaridade* (DMS). A Tabela 6.10 exhibe os resultados para valores de DMS iguais a 0,0 e 0,01.

Considerar uma DMS igual a zero significa que um documento só será classificado numa categoria se este obtiver similaridade maior ou igual à similaridade mínima calculada para o vetor desta categoria (*Sim-min*). Quando a DMS é maior que zero, é dada aí uma tolerância do tamanho da DMS, e documentos com similaridade abaixo da similaridade mínima calculada para o vetor da categoria podem ser categorizados nesta.

Tabela 6.10– Padrão de Erros da Categorização para DMS = 0,0 e 0,01

<i>Categoria que o documento deveria ser classificado</i>	<i>Categorias que predominantemente foram atribuídas erradamente ao documento</i>	
	<i>Até 300 termos</i>	<i>Acima de 300 termos</i>
CPOLI	Não identificada	ECONO ANTROP
ECONO	Não identificada	ESTPOP CPOLI
ESTPOP	Não identificada	ECONO CPOLI
ANTROP	Não identificada	ESTPOP ECONO CPOLI
FJN-OUTRAS (Não identificada)	ANTROP CPOLI	CPOLI ECONO ESTPOP

Observa-se na Tabela 6.10 que para até 300 termos processados, os erros de categorização das quatro categorias previstas pelo Modelo (CPOLI, ECONO, ESTPOP e ANTROP) se concentraram na categoria *Não identificada*. Isto ocorreu por causa da escassez de termos envolvidos no cálculo da similaridade (mesmo com tolerância de 0,01 oferecida pela DMS).

No caso de documentos da categoria FJN-OUTRAS (que o correto seria ser categorizado como categoria *Não identificada*), aqueles classificados erradamente caíram nas categorias ANTROP e CPOLI que, como visto na Tabela 6.2, foram as categorias que geraram os menores valores de similaridade mínima (*Sim-min*).

Quando os documentos passaram a ser processados considerando-se mais de 300 termos, o Modelo obteve melhores condições de avaliar a similaridade destes. Dentre os

documentos categorizados erradamente, alguns documentos da categoria FJN-CPOLI se identificaram predominantemente com a categoria ECONO e ANTROP, mas não com a categoria ESTPOP. Isto significa dizer que alguns documentos da categoria FJN-CPOLI continham palavras ditas significantes para as categorias ECONO e ANTROP, com alta frequência. A mesma analogia pode ser feita para os documentos das categorias FJN-ECONO e FJN-ESTPOP, concluindo-se que alguns documentos da categoria FJN-ECONO se identificaram predominantemente com a categoria ESTPOP e CPOLI, mas não com a categoria ANTROP, e alguns documentos da categoria FJN-ESTPOP se identificaram predominantemente com a categoria ECONO e CPOLI, mas não com a categoria ANTROP.

Chama-se também atenção para os documentos das categorias FJN-ANTROP e FJN-OUTRAS, que se identificaram com todas as outras categorias, mostrando a sua característica de assunto geral, não muito específico. Outra observação interessante é que os documentos da categoria FJN-OUTRAS não mais foram classificados na categoria ANTROP quando passaram a ser processados com mais de 300 palavras, concluindo-se que estes não trataram de assuntos muito parecidos, isto é, suas palavras mais frequentes não eram as mesmas da categoria ANTROP.

Para o caso de padrão de erro de categorização quando o valor da DMS foi igual a 0,02, a Tabela 6.11 exibe os seguintes resultados:

Tabela 6.11 – Padrão de Erros da Categorização para DMS = 0,02

<i>Categoria que o documento deveria ser classificado</i>	<i>Categorias que predominantemente foram atribuídas erradamente ao documento</i>	
	<i>Até 300 termos</i>	<i>Acima de 300 termos</i>
CPOLI	ANTROP	ECONO, ANTROP
ECONO	ANTROP	ESTPOP, CPOLI
ESTPOP	ANTROP	ECONO, CPOLI
ANTROP	CPOLI	ESTPOP, ECONO, CPOLI
FJN-OUTRAS (Não identificada)	ANTROP, CPOLI	CPOLI, ECONO, ESTPOP

Observa-se na Tabela 6.11 que para até 300 termos processados, os erros de categorização se concentraram nas categorias ANTROP e CPOLI. Isto é compreensível uma vez que a diferença máxima de similaridade (DMS) foi aumentada (0,02), fazendo com que

documentos com valores de similaridade abaixo da menor similaridade mínima se encaixassem em alguma categoria. Como as menores similaridades mínimas pertenciam às categorias ANTROP e CPOLI (ver Tabela 6.2), documentos recaíram nestas categorias.

Quando os documentos passaram a ser processados considerando-se mais de 300 termos, os erros de categorização foram semelhantes aos resultados da Tabela 6.10 (DMS igual a 0,0 e 0,1). Dessa forma, a análise feita anteriormente para esse item é válida também para a Tabela 6.11.

As próximas tabelas (6.12 a 6.17) exibem as taxas de acerto na geração das categorias dos documentos dos *corpora* para geração de metadados, para os algoritmos Porter e RSLP, quando se variou a quantidade de palavras a serem processadas de 100 até todas as palavras, e se fixou o parâmetro DMS para os valores 0,0; 0,01 e 0,02.

Tabela 6.12 – Taxas de Acerto na Categorização de Documentos (**RSLP, DMS = 0,0**)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	4	14,81%	13	48,15%	21	77,78%	26	96,30%	26	96,30%	26	96,30%
FJN-ECONO	21	1	4,76%	6	28,57%	9	42,86%	7	33,33%	7	33,33%	10	47,62%	13	61,90%
FJN-ESTPOP	20	0	0,0%	3	15%	6	30%	14	70%	18	90%	20	100%	20	100%
FJN-ANTROP	26	2	7,69%	5	19,23%	8	30,77%	13	50%	19	73,08%	19	73,08%	19	73,08%
FJN-OUTRAS	25	25	100%	22	88%	21	84%	18	72%	14	56%	11	44%	11	44%
Total	119	28	23,53%	40	33,61%	57	47,90%	73	61,34%	84	70,59%	86	72,27%	89	74,79%

Tabela 6.13 – Taxas de Acerto na Categorização de Documentos (**Porter, DMS = 0,0**)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	4	14,81%	14	51,85%	21	77,78%	26	96,30%	26	96,30%	27	100%
FJN-ECONO	21	0	0,0%	1	4,76%	4	19,05%	3	14,28%	9	42,88%	6	28,57%	12	57,14%
FJN-ESTPOP	20	0	0,0%	0	0,0%	1	5%	11	55%	18	90%	19	95%	20	100%
FJN-ANTROP	26	2	7,69%	5	19,23%	6	23,08%	12	46,15%	18	69,23%	19	73,08%	18	69,23%
FJN-OUTRAS	25	25	100%	22	88%	21	84%	20	80%	16	64%	14	56%	14	56%
Total	119	27	22,69%	32	26,89%	46	38,65%	67	56,30%	87	73,11%	84	70,59%	91	76,47%

Nas tabelas acima é interessante notar que para as categorias FJN-CPOLI, FJN-ECONO, FJN-ESTPOP e FJN-ANTROP, as taxas de acerto foram predominantemente crescentes, obtendo melhores resultados quando processados todos os termos dos documentos. Com pequena quantidade de termos (de 100 a 300), a escassez de palavras levou os documentos a serem classificados como categoria *Não identificada*, conforme visto na Tabela 6.10.

Para a categoria FJN-OUTRAS, seu comportamento foi o inverso, alcançando melhores resultados quando processados com 100 termos, e decaindo à medida que foi aumentando o número de termos. Observa-se, entretanto, que esses resultados foram decorrentes da escassez de palavras para o cálculo de similaridade, não se podendo concluir que a classificação *Não identificada* atribuída à categoria FJN-OUTRAS esteja correta para os primeiros 300 termos analisados.

Comparando-se ainda nestas tabelas os resultados obtidos para os *stemmer* RSLP e Porter (linhas de Total), verifica-se que o desempenho no geral foi ligeiramente melhor para o Porter, alcançando seu melhor resultado quando os documentos foram processados com todas as palavras (76,47% contra 74,79% para o RSLP).

As tabelas 6.14 e 6.15 abaixo exibem os resultados das taxas de acerto para os algoritmos RSLP e Porter respectivamente, quando a DMS foi igual a 0,01.

Tabela 6.14 – Taxas de Acerto na Categorização de Documentos (**RSLP, DMS = 0,01**)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	5	18,52%	13	48,15%	21	77,78%	26	96,30%	26	96,30%	26	96,30%
FJN-ECONO	21	1	4,76%	6	28,57%	9	42,86%	7	33,33%	7	33,33%	10	47,62%	13	61,90%
FJN-ESTPOP	20	0	0,0%	3	15%	6	30%	14	70%	18	90%	20	100%	20	100%
FJN-ANTROP	26	2	7,69%	5	19,23%	9	34,61%	14	53,85%	21	80,77%	21	80,77%	21	80,77%
FJN-OUTRAS	25	25	100%	22	88%	18	72%	14	56%	12	48%	10	40%	9	36%
Total	119	28	23,53%	41	34,45%	55	46,22%	70	58,82%	84	70,59%	87	73,11%	89	74,79%

Tabela 6.15 – Taxas de Acerto na Categorização de Documentos (**Porter, DMS = 0,01**)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	5	18,52%	14	51,85%	21	77,78%	26	96,30%	26	96,30%	27	100%
FJN-ECONO	21	0	0,0%	1	4,76%	4	19,05%	3	14,28%	9	42,88%	7	33,33%	12	57,14%
FJN-ESTPOP	20	0	0,0%	0	0,0%	1	5%	11	55%	18	90%	19	95%	20	100%
FJN-ANTROP	26	2	7,69%	5	19,23%	7	26,92%	14	53,85%	20	76,92%	22	84,61%	20	76,92%
FJN-OUTRAS	25	25	100%	22	88%	20	80%	18	72%	13	52%	11	44%	10	40%
Total	119	27	22,69%	33	27,73%	46	38,65%	67	56,30%	86	72,27%	85	71,43%	89	74,79%

Observando-se as tabelas acima, vê-se que seus resultados foram bem parecidos com os das tabelas 6.12 e 6.13. Da mesma forma que estas, para as categorias FJN-CPOLI, FJN-ECONO, FJN-ESTPOP e FJN-ANTROP, as taxas de acerto foram predominantemente crescentes, obtendo melhores resultados quando processados todos os termos dos documentos. Com pequena quantidade de termos (de 100 a 300), a escassez de palavras levou os documentos a serem classificados como categoria *Não identificada*, conforme visto na Tabela 6.10.

Para a categoria FJN-OUTRAS, seu comportamento foi o inverso, alcançando melhores resultados quando processados com 100 termos, e decaindo à medida que foi aumentando o número de termos. Observa-se, novamente, que esses resultados foram decorrentes da escassez de palavras para o cálculo de similaridade, não se podendo concluir que a classificação *Não identificada* atribuída à categoria FJN-OUTRAS esteja correta para os primeiros 300 termos analisados.

Comparando-se ainda nestas tabelas os resultados obtidos para os *stemmer* RSLP e Porter (linha de Total), verifica-se que suas taxas de acerto foram bem próximas, alcançando o melhor resultado (idêntico para os dois *stemmer*) quando os documentos foram processados com todas as palavras (74,79%).

As tabelas 6.16 e 6.17 abaixo exibem os resultados das taxas de acerto para os algoritmos RSLP e Porter respectivamente, quando a DMS foi igual a 0,02.

Tabela 6.16 – Taxas de Acerto na Categorização de Documentos (RSLP, DMS = 0,02)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	5	18,52%	13	48,15%	21	77,78%	26	96,30%	26	96,30%	27	96,30%
FJN-ECONO	21	1	4,76%	6	28,57%	9	42,86%	7	33,33%	7	33,33%	10	47,62%	13	61,90%
FJN-ESTPOP	20	0	0,0%	3	15%	6	30%	14	70%	18	90%	20	100%	20	100%
FJN-ANTROP	26	26	100%	26	100%	25	96,15%	23	88,46%	24	92,30%	23	88,46%	23	88,46%
FJN-OUTRAS	25	25	100%	19	76%	16	64%	5	20%	3	12%	2	8%	2	8%
Total	119	52	43,70%	59	49,58%	69	57,98%	70	58,82%	78	65,55%	81	68,07%	85	71,43%

Tabela 6.17 – Taxas de Acerto na Categorização de Documentos (Porter, DMS = 0,02)

<i>Categoria</i>	<i>Núm. Docs.</i>	<i>100 Termos</i>		<i>200 Termos</i>		<i>300 Termos</i>		<i>500 Termos</i>		<i>800 Termos</i>		<i>1200 Termos</i>		<i>Todos os Termos</i>	
		<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>	<i>N. docs corretos</i>	<i>Taxa acerto</i>
FJN-CPOLI	27	0	0,0%	5	18,52%	14	51,85%	22	81,48%	26	96,30%	26	96,30%	27	100%
FJN-ECONO	21	0	0,0%	1	4,76%	4	19,05%	3	14,28%	9	42,88%	7	33,33%	12	57,14%
FJN-ESTPOP	20	0	0,0%	0	0,0%	1	5%	11	55%	18	90%	19	95%	20	100%
FJN-ANTROP	26	26	100%	26	100%	25	96,15%	24	92,30%	24	92,30%	25	95,15%	24	92,30%
FJN-OUTRAS	25	25	100%	19	76%	17	68%	9	36 %	5	20%	4	16 %	4	16 %
Total	119	51	42,86%	51	42,86%	61	51,26%	69	57,98%	82	68,91%	81	68,07%	87	73,11%

Considerando-se agora uma DMS igual a 0,02, nota-se que apenas para as categorias FJN-CPOLI, FJN-ECONO e FJN-ESTPOP as taxas de acerto foram predominantemente crescentes, obtendo melhores resultados quando processados todos os termos dos documentos. Com pequena quantidade de termos (de 100 a 300), a escassez de palavras levou a maioria dos documentos dessas categorias a serem encaixados na categoria ANTROP, conforme visto na Tabela 6.11. Isto ocorreu porque a DMS ofereceu uma tolerância de 0,02 abaixo do valor de similaridade mínima oferecida pelo vetor de categoria. Como a menor similaridade mínima pertencia à categoria ANTROP (ver Tabela 6.2), documentos recaíram nesta categoria.

Para a categoria FJN-ANTROP, pela mesma explicação acima, seus documentos alcançaram 100% de taxa de acerto com até 200 termos processados por documento, decaindo levemente essa taxa à medida que foi aumentando o número de termos processados (chegando a 92,30% para o Porter). Destaca-se aqui que não se pode concluir que a taxa de acerto da categoria FJN-ANTROP, para os primeiros 300 termos processados por documento, esteja correta, já que ela foi obtida por escassez de termos para o cálculo da similaridade.

Para a categoria FJN-OUTRAS, a taxa de acerto de seus documentos também alcançou melhor resultado quando processados com 100 termos, decaindo à medida que foi aumentando o número de termos. Observa-se, novamente, que esses resultados foram decorrentes da escassez de palavras para o cálculo de similaridade, não se podendo concluir que a classificação *Não identificada* atribuída à categoria FJN-OUTRAS esteja correta para os primeiros 300 termos analisados.

Comparando-se ainda nestas tabelas os resultados obtidos para os *stemmer* RSLP e Porter (linhas de Total), verificou-se que o desempenho no geral foi ligeiramente melhor para o Porter, alcançando seu melhor resultado quando os documentos foram processados com todas as palavras (73,11% contra 71,43% para o RSLP).

As figuras 6.1 e 6.2 abaixo exibem os gráficos da taxa de acerto na categorização, por número de palavras processadas por documento, para a DMS igual a 0,0; 0,01 e 0,02. A variação do número de palavras começou a partir de 500 termos, uma vez que os resultados

até 300 termos processados não retrataram perfeitamente o processo de categorização (pela escassez de termos).

Figura 6.1 - Taxa de Acerto na Categorização dos *Corpora* para Geração dos Metadados – RSLP

Figura 6.2 - Taxa de Acerto na Categorização dos *Corpora* para Geração dos Metadados - Porter

Como dito anteriormente, de um modo geral o algoritmo Porter produziu uma maior taxa de acerto na categorização dos documentos dos *corpora* para geração dos metadados,

alcançando seu ponto máximo quando foram considerados todos os termos dos documentos e valor de DMS igual a zero (ver também Figura 6.3). De fato, à medida que se aumentou a DMS, os resultados foram declinando, mostrando que a tolerância oferecida por esta influenciou negativamente os resultados, categorizando os documentos em categorias erradas. Pode-se também concluir daí que os conjuntos de documentos que formaram cada uma das categorias continham documentos mínimos representativos.

A Figura 6.3 compara o desempenho do RSLP e do Porter para a taxa de acerto na categorização, quando foram processadas todas as palavras por documento.

Figura 6.3 - Taxa de Acerto na Categorização dos *Corpora* para Geração dos Metadados (Porter X RSLP)

Como forma de ampliar a avaliação dos resultados da categorização, observou-se o conteúdo de alguns documentos que não conseguiram ser categorizados corretamente. Abaixo se encontram pequenos trechos desses documentos, obtidos através do módulo de sumarização do Modelo. As palavras em negrito são aquelas que o Modelo utilizou para o cálculo da similaridade (palavras-chave).

- **Texto 1** - Documento pertencente à categoria FJN-ECONO classificado como FJN-CPOLI:

“Naquele mesmo ano, o **Sindicato** dos Servidores **Públicos** Federais, também filiado à CUT, **participou** realizando eleições e submetendo, sem êxito - diferentemente dos **jornalistas** -, três nomes ao **presidente** do TRT/PE. A situação em São **Paulo**, durante o **mandato** anterior dos representantes classistas naquele **Estado**, não era nada diferente (**Jornal da Tarde**, 2/9/91). O **presidente** daquela central, em Pernambuco, considerava, em 1992, necessário que seus **dirigentes** rediscutissem a posição de rejeitar, a indicação de candidatos “cutistas”. Aliás, a própria ANAMATRA - contrária à manutenção da representação classista - se opôs, na revisão constitucional, ao **controle** externo ao **Judiciário** (JC, 11/2/94). O próprio relator da revisão, o **deputado federal** Nélson Jobim, especulava com a hipótese de incluir no conjunto final das **propostas** da relatoria a extinção total da representação classista ou, pelo menos, sua manutenção apenas na primeira instância (JC, 13/3/94). Em Pernambuco, um ex - **presidente** do Tribunal Regional do **Trabalho** (TRT/PE) tem insistentemente afirmado que “se acabarem com o juiz classista, a Justiça do **Trabalho** fecha”, pois aquela representação é essencial à sua existência (JC, 18/7/93).”

- **Texto 2** - Documento pertencente à categoria FJN-ANTROP classificado como FJN-ESTPOP:

“É convencional situar o uso da ferramenta simples - ou objetos adaptados - nos **trabalhos** artesanais. Existe ainda o outro lado, massificante, que exige o cumprir **modelos** desvinculados das realidades locais. Tem - se aí apenas o labor, o **trabalho**. Daqueles **trabalhos** na área da **cultura**, não se **isola** o econômico, o político, o religioso, o **social**. Ou naqueles **trabalhos** na área da mão de obra, tem - se que respeitar a criação e o **modelo** da **comunidade**, pois o fazer não é **isolado** dos traços mais **identificadores** do **grupo**. O torno e a mesa do ceramista estão incorporados às conquistas técnicas e do **trabalho**. Os quadros ou grades das filezeiras, as almofadas de bilros das rendeiras, os bastidores do rendendê **integram** o elenco dos instrumentos de **trabalho** das rendas de agulha e de bilros. E a vida, é o **trabalho** de retirar do solo os alimentos, de marcar com rituais os plantios e as colheitas, nos rios, nos mares, na mineração, nos desmatamentos, no uso das encostas, da ocupação humana nos relevos, traçando seus comportamentos e dizendo, pelo **trabalho**, quem são, mantendo traços da **identidade**. O entendimento das materias - primas não está **isolado** na extração ou no tratamento do material”.

- **Texto 3** - Documento pertencente à categoria FJN-ANTROP classificado como FJN-ECONO:

“A literatura de cordel é assim chamada pela forma como são **vendidos** os folhetos, dependurados em barbantes (cordão), nas feiras, **mercados**, praças e bancas de jornal, principalmente das cidades do interior e nos subúrbios das grandes cidades. Os folhetos foram introduzidos no **Brasil** pelo cantador Silvino Pirauá de Lima e depois pela dupla Leandro Gomes de Barros e Francisco das Chagas Batista. O autor do folheto podia ficar num mesmo lugar a maior parte do tempo, porque suas obras eram **vendidas** por folheteiros ou revendedores empregados por ele. O poeta popular é o representante do povo, o repórter dos acontecimentos da vida no Nordeste do **Brasil**. Atualmente, a

literatura de cordel não tem um bom **mercado** no **Brasil**, como acontecia na década de 50, quando foram impressos e **vendidos** dois milhões de folhetos sobre a morte de Getúlio Vargas, num total de 60 títulos. Hoje, os folhetos podem ser encontrados em alguns **mercados** públicos, como o **Mercado** de São José, no Recife, em feiras, como a de Caruaru, e em sebos(**venda** de livros usados)”.

- **Texto 4** - Documento pertencente à categoria FJN-OUTROS classificado como FJN-CPOLI:

“Foi alfabetizado por um tio e dois amigos do seu pai. Antigos exemplares do **jornal** Diário de Pernambuco serviram - lhe de cartilha. Em 1908, ingressou na Faculdade de Direito do Recife indo **trabalhar** na época como aprendiz de repórter no **jornal** A Pátria. **Trabalhou** também no **Jornal** do Recife, no Diário de Pernambuco e no **Jornal** Pequeno, no qual publicou a maior parte de suas reportagens no Recife. Colaborou nos jornais A Época, **Jornal** do Commercio, Correio da Manhã, do Rio de Janeiro e também na edição vespertina d'O Estado de São Paulo. Seu sonho era adquirir um **jornal**, como primeiro elo de uma cadeia. Foi consultor para leis de guerra no Ministério das Relações Exteriores, no governo Nilo Peçanha, mas deixou o cargo para ser redator - chefe do **Jornal** do **Brasil**. Em 1919, depois de deixar o **Jornal** do **Brasil** foi convidado para ser correspondente internacional na Europa, **trabalhando** para o Correio da Manhã, do Rio de Janeiro. Viajou pela Suíça, Inglaterra, França, Holanda, Itália e Alemanha, obtendo sucesso **jornalístico** e pessoal. Em setembro de 1924, adquiriu O **Jornal**, do Rio de Janeiro, dando início à cadeia nacional de **jornal**, rádio e televisão dos Diários Associados, que iria revolucionar o **jornalismo brasileiro**, inovando a imprensa, modernizando equipes, processos e veículos. Em 1934, incorporou a sua cadeia o Diário de Pernambuco, o **jornal** mais antigo em circulação na América Latina, onde havia iniciado sua carreira de **jornalista**”.

Observando-se os textos acima e as listas de vetores de categorias disponíveis no Anexo B (Seção B.2), pode-se notar que realmente estes textos continham uma grande quantidade de palavras que caracterizavam a categoria que foram erradamente classificadas. Mesmo quando o vetor da categoria correta continha também essas palavras, por não serem as mais significativas para ela, seus pesos foram inferiores, gerando uma similaridade menor.

Por exemplo, o Texto 1 referenciado acima, que deveria ter sido categorizado como ECONO (mas recaiu na categoria CPOLI), gerou uma similaridade com ECONO de 0,079 e de 0,187 para CPOLI. Abaixo são exibidos alguns *stemming* e seus pesos produzidos para o vetor de palavras desse texto, para o vetor da categoria ECONO e da categoria CPOLI:

Tabela 6.18 – *Stemmings* e Pesos Produzidos para um Texto Exemplo

<i>Stemming</i>	<i>Peso no texto</i>	<i>Peso em ECONO</i>	<i>Peso em CPOLI</i>
control	0,10949536134961807	0,0021693669661075176	0,0064733090569885025
deput federal	0,06256877791406747	ausente	0,0035458234126942633
estad	0,2659173061347867	0,036445265519194164	0,11779853277711029
jornal	0,10949536134961807	ausente	0,003771958961544445
judiciári	0,06256877791406747	ausente	0,0026750571797725634
particip	0,07821097239258434	0,027180463974614197	0,005080068090429202
propost	0,07821097239258434	0,01551138571789555	0,005589548070082803
públic	0,06256877791406747	0,02042497468180376	0,008237493891712204
sindicat	0,17206413926368555	ausente	0,027166560083318716
trabalh	0,10949536134961808	0,0434025130241411	0,09040798609687557

6.2.2 Avaliação do Metadado *Description*

O metadado *description* foi obtido através da técnica de sumarização extrativa, que consiste em selecionar segmentos do texto fonte segundo alguns critérios pré-definidos, e justapô-los na mesma ordem em que são encontrados, compondo o sumário extrativo.

No experimento aqui apresentado, as regras adotadas para atribuição de pesos aos termos e seleção das sentenças foram àquelas definidas no Capítulo 5 (Seção 5.4.2). As métricas e métodos de avaliação também foram definidos no Capítulo 5 (Seção 5.5), sendo basicamente avaliados a coerência e a preservação da idéia central, além da capacidade de gerar categorias corretas. Todos esses elementos foram avaliados utilizando-se os algoritmos de *stemmer* RSLP e Porter, para diferentes taxas de compressão atribuídas aos sumários (70%, 80% e 90%).

A “coerência” do sumário e a preservação da idéia central foram avaliadas de forma humana. A coerência se baseou principalmente no critério de clareza de leitura do conteúdo e coesão entre as sentenças do sumário. A preservação da idéia central foi avaliada comparando-se o sumário com o texto fonte.

A Tabela 6.19 exhibe as características dos textos selecionados para fazerem parte desses experimentos. Como foi necessária a avaliação de juízes humanos, selecionaram-se os menores textos dos *corpora* para geração dos metadados.

Tabela 6.19 – Características dos Textos para Geração de Sumários Extrativos

<i>Categoria</i>	<i>Número de documentos</i>	<i>Tamanho em (KB)</i>	<i>Tamanho médio (KB) por documento</i>	<i>Número aproximado de palavras por documento</i>
Ciências Políticas	5	36	7,2	1.100
Economia	5	50	10	1.500
Estudos Populacionais	5	50	10	1.500
Antropologia Cultural	5	37	7,4	1.250
Total	20	173		

Para avaliação da coerência e da preservação da idéia central (aqui chamada de “qualidade do sumário”), adotaram-se os seguintes critérios e respectivos códigos:

Tabela 6.20 – Critérios para Atribuição de Qualidade aos Sumários

<i>Qualidade do sumário</i>	<i>Descrição</i>	<i>Código</i>
Bom	Sumário é coerente e preserva a idéia central	1
Razoável	Sumário não é coerente , mas preserva idéia central	2
Ruim	Sumário é coerente , mas não preserva idéia central	3
Péssimo	Sumário não é coerente e não preserva idéia central	4

Os textos selecionados foram submetidos ao *Módulo de Geração de Metadados* (Figura 5.3), e os sumários produzidos foram avaliados por três juízes humanos. Para se obter um resultado final único, considerou-se o resultado de menor qualidade atribuída pelos juízes (Tabela 6.21). O experimento foi executado processando-se todas as palavras de cada documento. Cada Juiz avaliou 140 documentos: 20 textos originais completos (de 3 a 5 páginas), e 6 sumários para cada texto original.

Tabela 6.21 – Avaliação Humana dos Sumários Extrativos

<i>Categoria</i>	<i>Texto</i>	<i>RSLP</i>			<i>Porter</i>		
		90%	80%	70%	90%	80%	70%
FJN-CPOLI	T1	1	2	1	1	2	1
	T2	2	2	1	2	2	1
	T3	4	2	2	2	1	1
	T4	2	1	1	2	2	1
	T5	2	2	2	2	2	2
FJN-ECONO	T1	2	2	2	2	2	2
	T2	2	2	2	4	2	1
	T3	4	2	2	2	2	2
	T4	1	2	2	4	1	1
	T5	1	1	1	4	2	2
FJN-ESTPOP	T1	2	2	1	2	1	2
	T2	2	1	2	2	1	2
	T3	2	2	2	2	1	1
	T4	2	2	2	2	2	2
	T5	2	2	2	2	2	2
FJN-ANTROP	T1	2	2	2	2	2	2
	T2	1	1	1	1	1	1
	T3	2	2	2	2	2	2
	T4	2	1	2	2	1	2
	T5	2	1	1	2	1	1

Observando-se a Tabela 6.21, nota-se que nem sempre os algoritmos RSLP e Porter produziram avaliações semelhantes, indicando a geração de sumários distintos.

Para efeito de ilustração, abaixo se encontram os sumários produzidos para o Texto T5 da categoria FJN-ECONO (90% de compressão), utilizando-se os algoritmos Porter e RSLP, e também para o Texto T3 da categoria FJN-CPOLI (90% de compressão). As frases sublinhadas foram àquelas escolhidas pelo Modelo como sendo as de maior peso (*gist sentence*). Os textos originais geradores desses sumários encontram-se no Anexo B (Seção B.4).

- **Texto T5** - Traços gerais do desenvolvimento recente do Nordeste. Categoria FJN-ECONO; Taxa de compressão 90%.

Resumo com *Stemmer* Porter (qualidade péssima):

Vetor de palavras-chave: ativ, comunic, crescent, dinâm, econom brasileir, econom nordestin, export, fbkf, indústr, long, méd, nord, país, produt, região, setor públic.

“Conseqüentemente, o PIB per capita do **Nordeste** em 1998 atingia apenas o patamar de 56,1% do respectivo PIB nacional (SUDENE, 1999). Além disso, a desvantagem no **produto** per capita é acrescida por uma distribuição de renda regional ainda pior que a **média** do **país**. É também verdade que esse peso maior do setor serviços resulta de um conjunto bastante heterogêneo de **atividades**, onde estão presentes níveis diferentes de **produtividade**, muitas vezes dentro do mesmo segmento. Em que pese esta menor abertura, há que se considerar que a participação dos **produtos** manufaturados e semimanufaturados nas **exportações nordestinas** vem crescendo. Com isso, agregou - se à pauta tradicional (açúcar principalmente) uma nova lista de **produtos exportáveis**. Entre estes encontram - se **produtos** petroquímicos (pólo de Camaçari), alumínio e uma gama diversificada de **produtos** manufaturados (confeções, **produtos** alimentícios, **produtos** do setor metal - mecânico, etc).”

Resumo com Stemmer RSLP (qualidade boa):

Vetor de palavras-chave: comunic, cresc, dinam, econom brasile, econom nordestin, export, fbkf, industr, long, med, pais, produt, region, set public.

“Nas últimas décadas a participação do PIB per capita do Nordeste no PIB per capita do Brasil **cresceu** de 41,8% para 56,1% entre 1960 e 1999. Em parte, pelo menos, isto resulta de um **crescimento econômico** relativamente mais rápido experimentado pela **economia nordestina** nas últimas quatro décadas (Tabela 1). O fato de ter **crescido** à frente da **economia brasileira** ao longo do período 1960 - 99, mostra essa capacidade de **crescimento**, tanto em períodos de alto quanto de baixo **crescimento** do conjunto da **economia nacional**. O **setor industrial**, que **cresceu** a 9,1% ao ano nos anos 1970, reduz abruptamente seu patamar de **crescimento** nos anos 1980/90 (1,0% ao ano), porém recupera parte do vigor nos anos 1990/99 (4,0% ao ano). Na indústria pode ser observado um padrão mais regular de comportamento, tendo os segmentos energia elétrica e abastecimento de água, em grupo ou individualmente, bem como a construção liderado os índices de **crescimento**. Destaque - se aqui o **crescimento** de 2, 1% ao ano observado no estado da Bahia, bastante superior, por exemplo, ao verificado em São Paulo(0, 8%) no mesmo período(SUDENE, 1999).”

- Texto T3 - Getúlio Vargas e o estado nacional desenvolvimentista. Categoria FJN-CPOLI, taxa de compressão 90%:

Resumo com Stemmer RSLP (qualidade péssima):

Vetor de palavras-chave: brasil, brasile, capit, consolid, cresc, desenvolviment, econom, est, industr, industri, mund, pais, polit.

“A produção **industrial cresce** então mais rapidamente, quase duas vezes, que a produção agrícola. na transferência de seus grandes grupos **industriais** pelo **mundo** e na **consolidação** de uma força militar que garante a aplicação da pax americana, exercida até hoje a partir da persuasão armada quando o comércio não é suficientemente capaz de impô-la. O **Brasil** terá que aguardar a passagem medíocre do governo Dutra na Presidência, para viver o período de profundo dinamismo de Vargas, com a definição e

consolidação da indústria de base, do petróleo, da energia e da química. Nos faltava a indústria de bens de consumo durável e a ampliação das indústrias de matérias primas e de máquinas.”

Resumo com Stemmer Porter (qualidade Razoável):

Vetor de palavras-chave: brasil, brasileir, consolid, desenvolviment, econôm, estad, estrutur, industrializ, indústr, país, polít.

“Quem quer que busque entender a política econômica brasileira de 1930 a 1964 não pode prescindir da figura de Getúlio Vargas, o presidente que mais tempo permaneceu à frente do governo brasileiro, seja na qualidade de líder revolucionário (1930 - 37), ditador (1937 - 45), durante o Estado Novo, e presidente constitucional (1950 - 54). Nos Estados Unidos da América, a política do New Deal, implantada por Roosevelt, foi uma política keynesiana antes de Keynes, cujo similar no Brasil foi a política de sustentação dos preços do café e sua queima posterior. A ideologia desenvolvimentista populista institucionalizava o poder político das forças populares, incorporando a emergência das massas no jogo político nacional.”

A Tabela 6.22 exibe um resumo dos resultados obtidos para a qualidade dos sumários produzidos. Os gráficos correspondentes aos dados dessa tabela estão nas figuras 6.4 e 6.5.

Tabela 6.22 – Qualidade dos Sumários Extrativos

Qualidade do Sumário	RSLP			Porter		
	90%	80%	70%	90%	80%	70%
Bom	4	6	7	2	8	9
Razoável	14	14	13	15	12	11
Ruim	0	0	0	0	0	0
Péssimo	2	0	0	3	0	0

Figura 6.4 – Qualidade dos Sumários Extrativos (RSLP)

Figura 6.5 – Qualidade dos Sumários Extrativos (Porter)

Como pode ser visto nos gráficos acima, de acordo com a avaliação dos juízes, o Modelo produziu, em sua grande maioria, extratos bons e razoáveis, preservando a idéia central do texto (100% para taxas de compressão iguais a 80% e 70%, independente do algoritmo de *stemmer* utilizado). Pode-se notar também que à medida que a taxa de compressão dos textos diminuiu (produzindo sumários maiores), a capacidade de gerar

sumários bons aumentou (chegando ao cume aos 70%), mostrando que muitos daqueles que apenas preservavam a idéia central do texto (extratos razoáveis) passaram também a ser coerentes (extratos bons). A razão para isso foi a regra de incorporar sentenças localizadas antes e depois das sentenças escolhidas, desde que dentro do mesmo parágrafo e respeitando-se a taxa de compressão (ver *sumarização* no Capítulo 5, seção 5.4.2).

Comparando-se o desempenho dos algoritmos Porter e RSLP, de um modo geral o Porter demonstrou uma melhor capacidade de gerar sumários bons que o RSLP, apesar de que para a taxa de compressão de 90% esse desempenho tenha sido um pouco inferior. As figuras 6.6 e 6.7 exibem a qualidade dos sumários produzidos com 70% de taxa de compressão, já que esta ofereceu os melhores resultados:

RSLP (Tx. de compressão = 70%)

Figura 6.6 – Qualidade dos Sumários Extrativos para Taxa de Compressão de 70% (RSLP)

Porter (Tx. de compressão = 70%)

Figura 6.7 – Qualidade dos Sumários Extrativos para Taxa de Compressão de 70% (Porter)

De acordo com os gráficos acima, vê-se que tanto o algoritmo RSLP como o Porter foram capazes de produzir extratos que preservam a idéia central do texto (bons e razoáveis), na sua totalidade (100%). Entretanto, para o item coerência, mais uma vez o algoritmo Porter se destacou, produzindo resultados superiores aos do algoritmo RSLP, oferecendo sumários de melhor qualidade (45% contra 35% do RSLP).

Dando continuidade à avaliação do metadado *description* produzido pelo Modelo, o experimento descrito abaixo teve o objetivo de verificar se os sumários extrativos do experimento anterior apresentaram informação suficiente para a correta categorização dos textos fonte.

Para tanto, os sumários produzidos pelo Modelo e seus respectivos textos originais (*corpus* para geração de sumários extrativos descrito na Tabela 6.19) foram submetidos ao *Módulo de Geração de Metadados*, e foi avaliada tanto a taxa de acerto na categorização, como também o tempo necessário para a realização da tarefa.

Este experimento foi realizado para os sumários produzidos com o algoritmo Porter, com taxa de compressão de 70%. Para a categorização, o parâmetro *diferença máxima de similaridade* (DMS) foi zero, e foram consideradas todas as palavras do documento. Estes parâmetros foram assim configurados porque os mesmos produziram os melhores resultados para a sumarização e para a categorização nos experimentos anteriores. A Tabela 6.23 exhibe os resultados desse experimento.

Tabela 6.23 – Taxas de Acerto na Categorização de Sumários Extrativos (Porter, DMS= 0,0; Taxa de compressão de 70%)

<i>Categoria</i>	<i>Textos originais</i>		<i>Sumários</i>	
	<i>Taxa acerto</i>	<i>Tempo process (seg)</i>	<i>Taxa acerto</i>	<i>Tempo process (seg)</i>
CPOLI	100%	3,0	80%	1,4
ECONO	100%	3,5	100%	2,0
ESTPOP	100%	3,0	80%	1,2
ANTROP	100%	2,5	80%	1,8
Média	100%	3,0	85%	1,6

Analisando-se a Tabela 6.23 se vê que cerca de 85% dos sumários foram bem sucedidos em sua categorização, sendo processados em aproximadamente 1,6 segundos (46,7% a menos que o tempo necessário para processar os textos originais).

Apesar de sua taxa de acerto ter sido 15% inferior à categorização dos textos originais, esta ainda é uma taxa muito alta, considerando-se que os sumários foram reduzidos a 70% de seus textos originais e o parâmetro de configuração *frequência mínima* dos termos de cada documento foi igual a três.

6.2.3 Avaliação dos Documentos RDF/XML

Uma vez extraídos os metadados dos documentos, estes foram convertidos para o formato RDF/XML, com *tags* que seguiram o padrão de metadados *Dublin Core*. Abaixo se encontra um exemplo de documento RDF/XML gerado automaticamente pelo Modelo. Para efeito de ilustração, o Anexo B (Seção B.5) apresenta alguns outros desses documentos.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<?xml-stylesheet href="modelo.xsl" type="text/xsl"?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xmlns:dcterms='http://purl.org/dc/terms/'>
<rdf:Description>
  <dc:subject>
    <rdf:Bag>
      <rdf:li>Antropologia</rdf:li>
      <rdf:li>african</rdf:li>
      <rdf:li>berimbau</rdf:li>
      <rdf:li>capoeira</rdf:li>
      <rdf:li>jog</rdf:li>
      <rdf:li>mestr</rdf:li>
    </rdf:Bag>
  </dc:subject>
  <dc:description>
    CAPOEIRA
    Desde os primeiros registros sobre a capoeira no Brasil, existem
    dúvidas sobre se sua origem veio de aldeias africanas ou de
    senzalas brasileiras. Os capoeiras, como eram chamados, não eram
    apenas escravos ou negros. Prostitutas, assaltantes, pedintes e
    ladrões de residências também faziam parte dos grupos de
    capoeiras. Durante as manifestações públicas e os períodos
    eleitorais, alguns políticos se aliavam aos capoeiras, que
    usavam sua força física para inibir a participação de defensores
    do partido adversário, em troca de benefícios. Apesar do seu
    aspecto violento, a capoeira também era uma forma de lazer,
    sendo acompanhada por instrumentos musicais. A capoeira, como
```


jogo ou dança, é executada no centro de um círculo de pessoas, a roda, que delimita o espaço onde ela se desenvolverá. O capoeirista tem que ser leve, flexível e ter muita ginga, pois é dela que partem os golpes de ataque e defesa. Seus movimentos são contínuos e a cadência é fundamental. Não se usa sapatos: luta - se ou joga - se a capoeira descalço. Hoje, alguns capoeiristas usam camisas coloridas com a logomarca da associação a qual pertencem. No Brasil, há grandes centros de capoeira na Bahia, em Pernambuco, no Rio de Janeiro e em São Paulo. Há uma corrente que defende uma nova modalidade de capoeira, a contemporânea, que é alvo de críticas por desprezar legados culturais do passado, mas para os que a defendem como espetáculo, competição e luta seria o caminho para chegar às Olimpíadas.

```
</dc:description>
<dc:publisher>FJN</dc:publisher>
<dc:date>2006-01-17</dc:date>
<dc:type>text/xml</dc:type>
<dc:format>text</dc:format>
<dc:identifier>ANTROP25</dc:identifier>
<dc:language>pt-br</dc:language>
<dc:rights>FJN</dc:rights>
<dc:title></dc:title>
<dc:author></dc:author>
<dc:contributor></dc:contributor>
<dc:relation></dc:relation>
<dc:source></dc:source>
<dc:coverage></dc:coverage>
</rdf:Description>
</rdf:RDF>
```

O documento acima mostra todos os metadados em formato DC gerados automaticamente pelo Modelo.

O metadado *subject* contém a categoria do texto e a lista de termos selecionados como significativos para ele (seus *stemmigs*). O metadado *description* contém o sumário extrativo. Os metadados *Rights*, *Format*, *Language*, *ObjectType* e *Publisher*, como dito no Capítulo 5, foram gerados com base nos parâmetros de usuários fornecidos pela interface do Módulo de Configuração (Figura 5.1).

O metadado *Identifier* foi composto pelo conteúdo do parâmetro *Identificador* fornecido pelo usuário no Módulo de Extração de Padrões de Categorias, concatenado a um número sequencial gerado automaticamente pelo Modelo para cada uma das categorias.

O metadado *Date* foi gerado a partir do sistema operacional. Para os elementos de metadados *Title*, *Author*, *Contributor*, *Relation*, *Source* e *Coverage*, o Modelo ofereceu sua geração semi-automática, fornecendo a sua codificação em RDF/XML com conteúdo vazio,

para serem posteriormente preenchidos de forma manual. Este procedimento semi-automático proporciona ao administrador do ambiente uma maior flexibilidade na complementação dos documentos digitais.

Como dito no capítulo anterior, cada documento RDF/XML foi validado por elementos disponibilizados pela Internet, invocados em suas primeiras linhas:

```
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:dc='http://purl.org/dc/elements/1.1/'
xmlns:dcterms='http://purl.org/dc/terms/'>
```

Para a formatação de exibição, um documento XSL (invocado na linha: `<?xml-stylesheet href="modelo.xsl" type="text/xsl"?>`) foi desenvolvido (ver Anexo B, Seção B.6).

Para avaliar os documentos RDF/XML, o Modelo contou com o aplicativo de RI e estratégia de avaliação definidos no Capítulo 5 (seção 5.5). Os *corpora* de textos que geraram documentos RDF/XML para serem avaliados foram os mesmos que compuseram os *Corpora* para Geração dos Metadados, nas categorias FJN-CPOLI, FJN-ECONO, FJN-ESTPOP, FJN-ANTROP (Tabela 6.24). Os parâmetros utilizados foram: taxa de compressão de 70%, DMS igual a zero, algoritmo Porter, processamento de todas as palavras. Estes parâmetros foram assim configurados porque os mesmos produziram os melhores resultados para a sumarização e para a categorização nos experimentos anteriores.

Tabela 6.24 – *Corpora* de Textos para Geração de Documentos RDF/XML

<i>Nome da Categoria</i>	<i>Sigla</i>	<i>Número de documentos</i>	<i>Tamanho em (KB)</i>	<i>Tamanho médio (KB) por documento</i>	<i>Número aproximado de palavras por documento</i>
Ciências Políticas	FJN-CPOLI	27	683	25,3	3.730
Economia	FJN-ECONO	21	607	28,9	4.300
Estudos Populacionais	FJN-ESTPOP	20	664	33,2	4.641
Antropologia Cultural	FJN-ANTROP	26	192	7,38	1.000
Total		94	2.146		

A Tabela 6.25 exhibe as características dos *corpora* de documentos RDF/XML, gerados a partir dos documentos dos *corpora* da Tabela 6.24.

Tabela 6.25 – Corpora de Documentos RDF/XML

<i>Número de documentos</i>	<i>Tamanho em (KB)</i>	<i>Tamanho médio (KB) por documento</i>	<i>Número aproximado de palavras por documento</i>
94	742	7,89	1.000

Para esse experimento, o aplicativo de RI foi executado por uma pessoa que desconhecia o conteúdo das bases textuais, tendo conhecimento apenas das categorias que as compunham. Esta pessoa realizou diversas buscas com palavras-chave de acordo com seu próprio interesse. O resultado desse experimento encontra-se na Tabela 6.26:

Tabela 6.26 – Resultados das Buscas com base nos Documentos RDF/XML e Documentos Texto

<i>Seq</i>	<i>Palavras-chave</i>	<i>Itens relevantes na coleção</i>	<i>RDF/XML</i>			<i>Texto</i>		
			<i>Tempo (milisseg)</i>	<i>Itens recup</i>	<i>Itens relev recup</i>	<i>Tempo (milisseg)</i>	<i>Itens recup</i>	<i>Itens relev recup</i>
1	reforma, agrária	7	20	5	5	2.003	13	7
2	golpe, estado	5	20	6	5	2.674	14	5
3	jk	2	10	2	2	2.674	2	2
4	dívida externa	1	10	1	1	3.876	2	1
5	fmi	1	10	2	1	3.355	4	1
6	eleições, pernambuco	3	10	3	2	3.195	9	3
7	partidos, políticos	11	110	22	11	3.455	15	9
8	candomblé	2	10	4	2	3.415	4	2
9	folclore, brasileiro	26	41	6	6	3.536	14	14
10	são, joão	3	40	10	3	2.153	29	3
11	carnaval	5	10	6	5	1.993	6	5
12	artesanato	3	30	3	3	1.792	5	3
13	dança, folclore	11	10	3	3	2.193	8	6
14	dança	11	45	25	8	2.824	61	9
15	êxodo, rural	1	10	1	1	2.033	6	1
16	movimento, sem, terra	5	125	13	5	3.125	26	5
17	mortalidade, envelhecimento	8	60	7	7	2.413	13	8
18	ptb	5	30	5	3	2.413	9	5
19	amamentação	2	10	2	2	2.724	3	2
20	renda, familiar	10	90	11	8	3.605	20	10
Total		122	701	137	83	55.451	263	101
Média		6,1	35,05	6,85	4,15	2.772,55	13,15	5,05

Na Tabela 6.26 destaca-se principalmente o tempo médio das buscas. Como era de se esperar, por serem bem menores, buscas em documentos RDF/XML foram realizadas em tempo expressivamente menor (cerca de 98,73%). A Tabela 6.27 exhibe os valores das métricas de *precisão* (número de itens relevantes recuperados dividido pelo número total de itens recuperados) e de *revocação* (número de itens relevantes recuperados dividido pelo número de itens relevantes na coleção). A precisão avalia se os documentos recuperados são todos relevantes. A revocação avalia se todos os documentos relevantes são recuperados.

Tabela 6.27 – Resultados das Buscas com base nos Documentos RDF/XML e Documentos Texto

<i>Seq</i>	<i>Palavras-chave</i>	<i>RDF/XML</i>		<i>Texto</i>	
		<i>Precisão</i>	<i>Revocação</i>	<i>Precisão</i>	<i>Revocação</i>
1	reforma, agrária	100,0%	71,4%	53,8%	100,0%
2	golpe, estado	83,3%	100,0%	35,7%	100,0%
3	jk	100,0%	100,0%	100,0%	100,0%
4	dívida externa	100,0%	100,0%	50,0%	100,0%
5	fmi	50,0%	100,0%	25,0%	100,0%
6	eleições, pernambuco	66,7%	66,7%	33,3%	100,0%
7	partido, político	50,0%	100,0%	60,0%	81,8%
8	candomblé	50,0%	100,0%	50,0%	100,0%
9	folclore, brasileiro	100,0%	23,1%	100,0%	53,8%
10	são, joão	30,0%	100,0%	10,0%	100,0%
11	carnaval	83,3%	100,0%	83,3%	100,0%
12	artesanato	100,0%	100,0%	60,0%	100,0%
13	dança, folclore	100,0%	27,3%	75,0%	54,5%
14	dança	32,0%	72,7%	14,7%	81,8%
15	êxodo, rural	100,0%	100,0%	16,7%	100,0%
16	movimento, sem, terra	38,5%	100,0%	19,2%	100,0%
17	mortalidade, envelhecimento	100,0%	87,5%	61,5%	100,0%
18	ptb	60,0%	60,0%	55,5%	100,0%
19	amamentação	100,0%	100,0%	66,7%	100,0%
20	renda, familiar	100,0%	71,4%	53,8%	100,0%
Média		60,6%	68,0%	38,4%	82,8%

Pela Tabela 6.27 se observa que as buscas baseadas em documentos RDF/XML ofereceram uma precisão média de 60,6%, o que correspondeu a 22,2% a mais de documentos recuperados relevantes que as buscas baseadas em documentos texto. Com relação a se todos os documentos relevantes foram recuperados (revocação), sua taxa foi

14,8% inferior à média alcançada pelas buscas em documentos texto. Apesar disso, juntando-se a esses resultados o tempo expressivamente menor obtido pelas buscas baseadas em documentos RDF/XML (cerca de 98,73% a menos), conclui-se que o Modelo ofereceu resultados promissores, mostrando que a *representação e a indexação de documentos baseados no contexto das informações* participam de forma fundamental no desempenho de sistemas de informação.

É importante lembrar aqui que um dos maiores problemas da *Web* é a tendência dos mecanismos de busca de inundarem os usuários com resultados irrelevantes (baixa precisão), por não levarem em consideração o contexto específico desejado, exigindo que usuários analisem cada uma das informações individualmente, e descartem aquelas (geralmente maioria) que não são de seus interesses.

As figuras abaixo exibem a comparação das medidas de precisão e revocação, respectivamente.

Figura 6.8 – Medidas de Precisão obtidas em Buscas a Documentos RDF/XML e Documentos Texto

Figura 6.9 – Medidas de Revocação obtidas em Buscas a Documentos RDF/XML e Documentos Texto

Conclusões e Trabalhos Futuros

Esta Tese ofereceu uma contribuição na área de *Web Semântica*, no âmbito da *representação e indexação de documentos*, definindo o SIGAMeta, um *Modelo de geração automática de metadados baseado em contexto, a partir de documentos textuais na língua portuguesa, em formato não estruturado* (txt). Um conjunto teórico amplo de assuntos ligados à criação de ambientes digitais semântico também foi apresentado.

Conforme recomendado em SemanticWeb.org, os documentos textuais aqui estudados foram automaticamente convertidos em páginas *Web* anotadas semanticamente, utilizando o *Dublin Core* como padrão para definição dos elementos de metadados, e o padrão RDF/XML para representação dos documentos e descrição dos elementos de metadados.

Dentre os quinze elementos de metadados *Dublin Core*, nove foram gerados automaticamente pelo Modelo, e seis foram gerados de forma semi-automática. A geração semi-automática proporciona ao administrador de um ambiente uma maior flexibilidade na complementação de documentos digitais.

A escolha dos elementos do DC para geração automática foi baseada nos fatores de “importância na retenção de contexto”, “possibilidade de extração do elemento a partir do conteúdo do documento” e “possibilidade de extração do elemento a partir de parâmetros fornecidos pelo usuário ou pelo sistema operacional”.

Dentre os elementos DC gerados automaticamente, *Description* e *Subject* foram os que necessitaram de algoritmos mais complexos, sendo obtidos através de técnicas estatísticas, de mineração de textos e de processamento de linguagem natural.

Para a realização dos experimentos, o Modelo contou com dois conjuntos de documentos textuais: *corpora de treinamento* e *corpora para geração de metadados*.

Os *corpora* de treinamento foram compostos por 208 textos científicos pertencentes às categorias Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural. Estes foram recolhidos pela Internet, em *sites* especializados nas categorias.

Os *corpora para geração dos metadados* foram compostos por 119 textos e, igualmente aos *corpora* de treinamento, foram formados pelas categorias Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural. Foram também considerados documentos pertencentes a outras categorias, objetivando verificar o comportamento do Modelo diante de categorias não previstas. Todos esses documentos foram recolhidos do *site* da Fundação Joaquim Nabuco (www.fundaj.gov.br), nas opções *Textos on-line* e *Pesquisa Escolar*.

O Modelo foi implementado em linguagem Java, e foram implementados algoritmos e métodos para obtenção de cada um dos metadados, para converter os documentos textuais em páginas *Web* anotadas semanticamente e para auxiliar na avaliação dos metadados produzidos.

O Modelo foi concebido para ser executado em duas fases: Fase de treinamento e Fase de geração dos metadados. Em ambas as fases, parâmetros de usuário foram coletados.

A Fase de treinamento teve o objetivo de extrair, a partir dos *corpora* de treinamento, os elementos que representaram cada categoria. A Fase de geração dos metadados executou os algoritmos que extraíram, para cada documento, cada um dos nove elementos de metadados do DC gerados automaticamente, a partir dos *corpora* para geração dos metadados e dos parâmetros de usuários. Uma vez extraídos os metadados, o Modelo converteu os documentos para o formato RDF/XML, complementando-os com os metadados a serem gerados de forma semi-automática.

A finalidade principal da avaliação do Modelo foi verificar o comportamento dos documentos convertidos para o formato RDF/XML, quando estes foram submetidos a um processo de recuperação de informação. Os elementos de metadados *Description* e *Subject* foram exaustivamente avaliados, uma vez que estes são os principais responsáveis por apreender a semântica de documentos textuais.

Apesar de não ser objetivo principal dessa Tese a avaliação de desempenho dos algoritmos de *stemmer* Porter e RSLP, estes foram avaliados em termos de influência na geração dos metadados *Description* e *Subject*. Vale a pena ressaltar que não se encontrou

trabalhos que tenham realizado avaliações de algoritmos de *stemmer* no âmbito da categorização e da sumarização de textos.

7.1 Conclusões

- Avaliação do Metadado *Subject*

O metadado *Subject* foi obtido através da técnica de categorização assistida, onde primeiramente o Modelo foi “ensinado” a identificar características específicas das categorias Ciências Políticas, Economia, Estudos Populacionais e Antropologia Cultural. Os textos dos *corpora* para geração dos metadados foram processados com os *stemmer* Porter e RSLP, e foi avaliada a taxa de acerto na geração das categorias dos documentos. A categorização foi atribuída com base na estratégia de *Categorização por Diferença de Similaridade*, definida nesta Tese. Esse experimento foi realizado variando-se a quantidade de palavras a serem processadas por documento, quando o parâmetro *diferença máxima de similaridade* – DMS assumiu valores 0,0; 0,01 e 0,02.

Dentre todos os resultados, a categorização obteve uma melhor taxa de acerto quando foram considerados todos os termos dos documentos e valor de DMS igual a zero (76,47% para Porter e 74,79% para RSLP). À medida que se aumentou a DMS (ou que se diminuiu o número de palavras processadas por documento), a taxa de acerto foi declinando, mostrando que a tolerância oferecida por esta influenciou negativamente os resultados, categorizando os documentos em categorias erradas.

Observando-se os textos categorizados de forma errada, notou-se que realmente estes continham uma grande quantidade de palavras que caracterizavam a categoria que foram erradamente classificadas. Mesmo quando o vetor da categoria correta continha também essas palavras, por não serem as mais significativas para ela, seus pesos foram inferiores, gerando uma similaridade menor.

- Avaliação do Metadado *Description*

O metadado *Description* foi obtido através de técnicas de sumarização extrativa de textos, sendo sua avaliação realizada pela verificação dos seguintes elementos: coerência (traduzida pela clareza de sua leitura), preservação da idéia central e capacidade de gerar

categorias corretas. Todos esses elementos foram avaliados para diferentes taxas de compressão (70%, 80% e 90%), e foram consideradas todas as palavras de cada documento.

Para avaliação da coerência e da preservação da idéia central (qualidade do sumário), documentos foram analisados por três juízes humanos, que os qualificaram como “bom”, “razoável”, “ruim” e “péssimo”. Para se obter um resultado final único, considerou-se o resultado de menor qualidade atribuída pelos juízes.

De acordo com suas avaliações, o Modelo produziu, em sua grande maioria, extratos bons e razoáveis, preservando a idéia central do texto (100% para taxas de compressão iguais a 80% e 70%, independente do algoritmo de *stemmer* utilizado). À medida que a taxa de compressão dos textos diminuiu (produzindo sumários maiores), a capacidade de gerar sumários bons aumentou (chegando ao cume com 70% de compressão), mostrando que muitos daqueles que apenas preservavam a idéia central do texto (extratos razoáveis) passaram também a serem coerentes (extratos bons). A razão para isso foi a regra de incorporar frases localizadas antes e depois das sentenças escolhidas, desde que dentro do mesmo parágrafo e respeitando-se a taxa de compressão.

Para o elemento “capacidade de gerar categorias corretas”, o objetivo foi verificar se os sumários possuíam informação suficiente para expressar a categoria do documento. Para tanto, um conjunto de sumários gerados antecipadamente pelo Modelo (taxa de 70% de compressão, algoritmo Porter e DMS igual a zero) foi submetido ao módulo de categorização, e avaliados a taxa de acerto e o tempo de categorização.

Uma média de 85% dos sumários foram bem sucedidos em sua categorização, sendo processados em aproximadamente 1,6 segundos (46,7% a menos que o tempo necessário para processar os textos originais). Apesar de sua taxa de acerto ter sido 15% inferior à categorização dos textos originais, esta ainda é uma taxa muito alta, considerando-se que os sumários foram reduzidos a 70% de seus textos originais e o parâmetro de configuração *frequência mínima* dos termos de cada documento foi igual a três.

- Avaliação de Desempenho dos Algoritmos de *Stemmer*

Na comparação de algoritmos de *stemmer*, tradicionalmente os experimentos medem a efetividade de desempenho baseada no percentual de redução de palavras a seu radical.

Nesta Tese, além de aplicar este tipo tradicional de avaliação, os algoritmos Porter e RSLP também foram avaliados quanto à qualidade de seus radicais e quanto ao seu desempenho perante resultados de categorização e sumarização dos textos.

Na análise da taxa de redução de palavras a seu radical, foram observados os termos gerados a partir dos *corpora* de treinamento, e contabilizados o número de palavras distintas antes do processo de *stemmer* e o número de *stemmings* gerados por estas palavras para o Porter e o RSLP.

Neste caso, o desempenho geral do algoritmo RSLP foi superior ao do algoritmo Porter, com taxa média de redução respectivamente de 46,22% e 39,64%. Entretanto, com relação à qualidade dos *stemmings* gerados, verificou-se que o algoritmo RSLP cometeu erros de *overstemming* (remoção de caracteres que na realidade não fazem parte do sufixo, e sim do radical), e o algoritmo Porter cometeu erros de *understemming* (não remoção completa do sufixo). Isto justifica o fato do algoritmo RSLP ter gerado uma menor quantidade de *stemmings* que o algoritmo Porter.

Considerando que erros de *overstemming* podem ser bem mais prejudiciais que erros de *understemming*, quando se tratando de aplicações que buscam extrair o contexto de informações, conclui-se que o algoritmo Porter obteve um melhor desempenho na geração de *stemmings* de melhor qualidade que o algoritmo RSLP.

No processo de *Categorização por Diferença de Similaridade*, os resultados obtidos com o *stemmer* Porter foi ligeiramente melhor que o RSLP, alcançando seu melhor resultado quando os documentos foram processados com todas as palavras e DMS igual a zero (76,47% contra 74,79% para o RSLP).

No processo de sumarização, tanto o algoritmo RSLP como o Porter foi capaz de produzir extratos que preservaram a idéia central do texto, na sua totalidade (100% para taxas de compressão iguais a 70% e 80%). Entretanto, para o item coerência, mais uma vez o algoritmo Porter se destacou, produzindo resultados superiores ao algoritmo RSLP, oferecendo sumários de melhor qualidade (45% contra 35% do RSLP).

- Avaliação dos Documentos RDF/XML

Nesta avaliação foram utilizadas as métricas de *precisão* e *revocação* para comparar o desempenho de buscas em base de dados de documentos texto (2.146 KB) e em base de

dados de documentos RDF/XML gerados pelo Modelo (742 KB), a partir de 94 documentos dos *Corpora* para Geração dos Metadados. Foi também investigado o tempo de resposta de cada caso.

Para esse experimento, um aplicativo de RI desenvolvido e executado por uma pessoa que desconhecia o conteúdo das bases de dados, tendo conhecimento apenas das categorias que as compunham. Esta pessoa realizou diversas buscas com palavras-chave de acordo com seu próprio interesse.

Como era de se esperar, por ser bem menor, buscas na base de dados de documentos RDF/XML foram realizadas em tempo expressivamente menor: cerca de 98,73% a menos que o tempo utilizado nas buscas em documentos texto.

As buscas baseadas em documentos RDF/XML ofereceram uma precisão média de 60,6%, o que correspondeu a 22,2% a mais de documentos relevantes recuperados que as buscas baseadas em documentos texto. Com relação a se todos os documentos relevantes foram recuperados (revocação), sua taxa foi 14,8% inferior à média alcançada pelas buscas em documentos texto. Apesar disso, juntando-se a esses resultados o tempo expressivamente menor obtido pelas buscas baseadas em documentos RDF/XML, conclui-se que o Modelo ofereceu resultados promissores, mostrando que a *representação e a indexação de documentos baseados no contexto das informações* participam de forma fundamental no desempenho de sistemas de informação.

Finalizando, apesar das técnicas aqui utilizadas na exploração do conteúdo dos documentos não terem sido verdadeiramente novas, não se pode ignorar que os elementos inovadores introduzidos pela Web Semântica e por conceitos e técnicas automáticas de potencialidade ofereceram avanços que possibilitaram a obtenção de resultados importantes nesta Tese.

Como demonstrado aqui, a junção dessas técnicas com os padrões e tecnologias recomendados pela Web Semântica pode minimizar um dos maiores problemas da *Web* atual, e uma das fortes razões para a implementação da Web Semântica: a tendência dos mecanismos de busca de inundarem os usuários com resultados irrelevantes, por não levarem em consideração o contexto específico desejado pelo usuário.

Dessa forma, é importante que se dê continuidade aos estudos e pesquisas em todas as áreas relacionadas à implementação da Web Semântica, dando abertura para que sistemas de informação mais funcionais sejam projetados.

7.2 Limitações

O Modelo aqui estudado dedicou-se ao processamento de documentos textuais não estruturados, na língua portuguesa. Entretanto, é possível a sua utilização para documentos em outros idiomas, adaptando-se a rotina de pré-processamento com dicionário e *stopwords* de outras línguas, ou utilizando novas rotinas de *stemmer* adaptadas à nova realidade. Além do mais, por fazer uso de métodos estatísticos e de PLN, este Modelo pode ser praticamente aplicado para textos de qualquer gênero e domínio, bastando que se treine antecipadamente o sistema com novos *corpora* específicos.

Apesar dos experimentos terem sido realizados com base em alguns gêneros e domínios específicos de conteúdo textual, essa limitação não invalida os resultados aqui obtidos, vindo a incentivar o desenvolvimento de novos trabalhos, onde alguns deles são sugeridos na próxima seção.

7.3 Trabalhos Futuros

A diversidade de contextos, a complexidade dos problemas relativos à língua portuguesa, e os novos conceitos introduzidos pelos padrões e tecnologias da Web Semântica foram alguns dos fortes desafios enfrentados na construção do Modelo aqui proposto. Entretanto, resultados promissores foram alcançados, incentivando investigações com o uso de técnicas e métodos mais avançados para o tratamento semântico de textos.

O Modelo aqui estudado pode ser complementado com outras opções de padrões de metadados, técnicas de medição de similaridade, cálculos de pesos, algoritmos de *stemmer* e rotinas que extraíam outros elementos, como por exemplo, o título e autor do documento. Base de dados contendo sinônimos também pode facilmente ser inserida.

Outras investigações podem ser realizadas com *corpora* de outros gêneros e domínios, e parâmetros de valores diferentes, para que resultados sejam comparados.

A geração de vetores representativos de categorias é fundamental na identificação de domínios textuais, exigindo a participação de especialistas das diversas áreas. A simples seleção de palavras baseadas em procedimentos estatísticos e técnicas de PLN não bastam. Assim, fica aí mais um desafio para a área de Inteligência Artificial: extrair conhecimento de especialistas das diversas áreas e convertê-lo em formato compreensível por máquina, auxiliando a geração de vetores específicos de categorias.

Apêndice A

Desafios na Construção e Manutenção de um Ambiente Digital Semântico

Este Apêndice apresenta alguns dos principais desafios enfrentados no desenvolvimento de ambientes digitais voltados para instituições centradas em gestão e disseminação da informação.

A.1 Desafios na Provisão de Acessos Efetivos:

a) Infra-estrutura tecnológica de suporte.

A formação de um ambiente digital deve contar com uma infra-estrutura tecnológica composta por diversos elementos, como por exemplo [ITESM1999]:

- Rede local de alta velocidade e conexões rápidas para a Internet;
- Conexões rápidas a outras instituições (museus, escolas, bibliotecas, ambientes culturais etc.), gerando pontos de presença;
- Servidores para gerenciamento, processamento, disponibilização de informação, armazenamento para diversos tipos de formatos digitais etc;
- *Hardware* para digitalização (gravadores de CD/DVD, estações de trabalho, *scanners*, câmera e filmadoras digitais, placas de captura de áudio e vídeo, digitalizadores 3D etc.);
- *Software* para administração; conversão para formatos digitais; edição e manipulação de texto, áudio e vídeo; geração, gerenciamento e manutenção de metadados; indexação e busca eletrônica; entrega e visualização; controle de acesso com segurança; gerenciamento de propriedade intelectual; replicação de repositórios; tarifação etc;
- Interfaces e protocolos para operação e integração com outros sistemas;
- Suportes para armazenamento de mídias digitais (fitas, CDs, DVDs etc.).

b) Obtenção de metadados que atendam efetivamente às necessidades de seus usuários.

Segundo Tannenbaum [Tannenb2001], para que metadados atendam efetivamente às necessidades de seus usuários, estes deveriam ser avaliados com relação às seguintes características definidas abaixo:

- A *precisão* de um metadado é avaliada em termos de se verificar se ele descreve seu objeto alvo de tal forma a nunca haver dúvidas sobre o que está sendo descrito;
- Um metadado é dito *consistente* se não houver nenhum outro que represente exatamente o objeto alvo que ele está representando;
- Um metadado se encontra em sua forma *modular* quando não pode mais ser quebrado em componentes;
- Todas as *perspectivas* de usuários devem estar representadas, contudo relacionadas a um elemento equivalente comum. Por exemplo, “contatos externos” pode ser referenciado para alguns usuários como “clientes”, e para outros como “usuários”. Tais perspectivas podem vir a se relacionarem pelo rótulo “contato”, por exemplo;
- *Identificação e categorização* de metadados necessitam seguir esquemas consistentes, de modo a oferecer facilidade em buscas e descobertas.

Os metadados de um ambiente digital semântico devem atender às características definidas acima para que possam oferecer acessos efetivos aos seus objetos de informação. Dois pontos importantes aqui se destacam: sincronização e padronização. Se metadados não são gerados no momento da criação de seu objeto de informação, a probabilidade destes se tornarem obsoletos ou fora de sincronização é grande. Adotar regras de padronização para atribuição de valores, identificadores, categorias e características aos metadados também é necessário. Se sincronismo e padronização na criação de metadados não podem ser garantidos, então sua precisão também não poderá ser garantida.

A tarefa de proporcionar acessos efetivos afetará as atividades internas da organização provedora do ambiente digital, exigindo-lhe a execução de procedimentos especiais para criação, manutenção e gerenciamento de seus metadados, garantindo-lhes a continuidade das características fundamentais definidas acima.

c) Gerenciamento, manutenção e disponibilização de metadados.

Tão importante quanto à avaliação dos metadados de um ambiente digital semântico, é a avaliação da metodologia adotada em seu gerenciamento, manutenção e disponibilização, uma vez que estes tenham sido criados. Deve-se acompanhar o fluxo de evolução de cada metadado em seu tempo de vida, para verificar se estes estão de acordo com as atualizações de seus respectivos objetos de informação, mantendo-se sincronizados [Gilliland2000]. As seguintes necessidades devem ser atendidas pela metodologia adotada:

- Criação, acesso e atualização dos metadados pelos usuários técnicos;
- Apresentação de metadados a usuários finais;
- Distribuição dos metadados tanto dentro como fora do ambiente digital;
- Disponibilização das especificações de metadados;
- Definição e cumprimento de regras de segurança dos metadados;
- Manutenção, adaptação, customização e expansão do ambiente digital.

d) Ferramentas que incorporam contribuições de usuários no sistema, gerando ambientes escaláveis, robustos, customizáveis, dinâmicos e úteis para diferentes comunidades e propósitos.

Com o objetivo de melhor aproveitar os recursos computacionais comumente escassos, na sustentação de ambientes digitais geralmente é necessário se reduzir o nível de detalhe oferecido nas descrições e catalogações de objetos de informação. Entretanto, tal procedimento conduz a objetos com informações incompletas, dificultando buscas e descobertas efetivas [Congress1].

Uma forma de aumentar o valor da informação, e conseqüentemente melhorar as chances de sucesso em buscas e descobertas, seria a incorporação do conceito de *usabilidade*. Usabilidade significa captar dos usuários do sistema informações que venham enriquecer a descrição ou catalogação dos objetos, como também informações que reflitam comportamentos, interesses e características próprias, contribuindo para a geração de ambientes digitais escaláveis, robustos, customizáveis, dinâmicos e úteis para diferentes comunidades e para diferentes propósitos.

Um ambiente digital semântico deve buscar soluções para tirar proveito dessas oportunidades, desenvolvendo ferramentas colaborativas que permitam grupos afins

proverem aumento de valor ao material buscado, de forma compartilhada, sem, entretanto, violar a privacidade desses, e se precavendo também contra o mau uso.

- e) Prover ferramentas mais eficientes e mais flexíveis para tornar o conteúdo digital adaptado às necessidades dos usuários finais, com relação a formatos e versões diferenciados.

Em ambientes digitais, é comum que objetos de informação se encontrem em formatos múltiplos ou versões, com o objetivo de atender às diversas necessidades de usuários, como preservação, pesquisa e disseminação. Objetos de informação podem ser reproduzidos em versões de tamanhos diferentes; versões com ou sem compressão de imagens, textos, áudio, e vídeo; textos formatados para serem exibidos em *browsers*, bem como para preservação ou publicação; versões tanto em formato proprietário, como em formato público ou “aberto”.

Os formatos digitais para armazenamento de objetos podem fornecer a base para assegurar a longevidade das informações e aperfeiçoamento de buscas e descobertas em bases de dados. A escolha de tais formatos influenciará no modo que o objeto de informação poderá ser explorado [Hodge2000]. Por exemplo, apenas alguns formatos oferecem pesquisa *full-text* em objetos de informação textual. O processo de digitalização para geração desses formatos é mais demorado e trabalhoso, por exigirem reconhecimento óptico dos caracteres (OCR), além de edição e manipulação do texto das obras. Entretanto, tais formatos costumam gerar arquivos de tamanhos menores, e conseqüentemente com velocidades de transmissão também menores.

Para a seleção de padrões de formatos digitais adequados aos diversos tipos de mídia, as seguintes características deveriam ser observadas [ITESM1999]:

- Ser amplamente utilizado no mercado;
- Ser um formato de padrão aberto ou suportado por produtos diversos;
- Ter um futuro claro em sua evolução;
- Interoperar com as várias tecnologias já existentes na organização (*softwares* de indexação e pesquisa, *browser* etc.);
- Oferecer uma qualidade de imagem (ou áudio) adequada;

- Atender às necessidades de exploração do objeto por parte do usuário (pesquisas *full-text* ou não);
- Levar em consideração os custos e benefícios advindos do processo de digitalização para geração daquele formato;
- Levar em consideração o tamanho de arquivo gerado;
- Levar em consideração a velocidade de transmissão de dados armazenados naquele formato.

Atualmente, devido às dificuldades tecnológicas de conversões de formatos e geração de versões de forma rápida, não é possível se manter uma única versão e formato de cada objeto de informação, sem comprometer a efetividade no atendimento das necessidades dos usuários. Dessa forma, é desejável que se criem meios efetivos de controle das diversas versões e formatos dos objetos de informação disponíveis no ambiente digital, possibilitando o sistema oferecer conteúdos digitais melhor adaptados às necessidades de preservação, pesquisa e disseminação de objetos de informação.

- f) Desenvolver propostas que possam recuperar e apresentar recursos heterogêneos de forma rápida, efetiva e completa.

Ferramentas automatizadas para busca e descoberta de informação digital têm utilizado, em sua grande maioria, indexação baseada em palavras-chave. Observa-se, entretanto, que apesar de tal indexação oferecer facilidades na localização de objetos de informação que contêm palavras-chave específicas, elas encontram dificuldades em simultaneamente identificar elementos que compartilham o mesmo contexto, isto é, pertencentes às mesmas classes de assuntos [JJBW1998-a]. Outro assunto importante é que os atuais avanços em tecnologias de digitalização e de *browse* têm aumentado a qualidade da apresentação dos objetos de informação, e conseqüentemente exigindo mecanismos de exibição e acesso também mais sofisticados.

Dessa forma, ambientes digitais semânticos necessitam ser dinâmicos e abertos a mudanças, para que possam absorver, sem grandes transtornos, toda essa diversidade de características de formatos de objetos, além de evoluções em mecanismos de busca, descoberta e exibição de elementos digitais. É desejável que se desenvolvam mecanismos

que permitam acesso aos elementos pesquisados de forma rápida, efetiva e completa, independente da heterogeneidade de características dos elementos pesquisados. Para tanto, é importante que todos os dados disponíveis e úteis sobre objetos de informação sejam usados na recuperação e busca de objetos, ao invés de forçar usuários a confiar em algum subconjunto da informação disponível, que poderá lhe trazer resultados fora do contexto desejado ou incompletos.

A.2 Desafios na Construção e Manutenção das Coleções Digitais:

a) Escolha de métodos para a composição das coleções digitais.

Um dos principais assuntos sobre criação de ambiente digital diz respeito à construção de suas coleções, a qual pode ser realizada por basicamente três métodos distintos [Cleve1998]:

- *Digitalização*, onde a informação representada por um objeto tangível ou por uma entidade analógica ou digital, é convertida para um objeto que possui um formato digital padrão;
- *Aquisição de trabalhos originais digitais*, como artigos, imagens e base de dados eletrônicos;
- *Acesso remoto a materiais externos* através da provisão de ponteiros para *Web sites* e outros ambientes digitais, mediante ou não acordo prévio entre instituições. Apesar deste método não exatamente constituir parte de um ambiente digital local, ainda é uma forma de aumentar os materiais disponíveis para os usuários.

A escolha de qual método (ou métodos) adotar influenciará no modo como a organização poderá gerenciar, oferecer acesso e preservar suas coleções. Produção e manutenção de versões múltiplas e formatos diferenciados de um mesmo objeto de informação também fazem parte da criação de um ambiente digital. Tais atividades são importantes para tornar o conteúdo digital adaptado às necessidades dos usuários, como preservação, pesquisa e disseminação.

Observa-se que práticas de digitalização e aquisição de trabalhos digitais são extremamente caras (os dois primeiros métodos), especialmente quando não se conta com

parceiros que possam se ajudar mutuamente. Ao trabalhar em conjunto, instituições com objetivos comuns podem ganhar maior eficiência e reduzir os custos globais envolvidos nesta atividade. Além do mais, também podem ser reduzidos a redundância e o desperdício da aquisição ou conversão de materiais duplicados. E finalmente, a construção de coleções digitais em conjunto traz o aumento do compartilhamento dos recursos e da riqueza das coleções que os usuários terão acesso.

Entretanto, o desafio aqui está em assegurar que tais coleções compartilhadas poderão passar por processos que as gerenciem, preservem e disponibilizem adequadamente por toda a sua vida útil. Por exemplo, se as únicas cópias dos trabalhos digitais residirem no servidor de uma determinada instituição, e se esta for desativada, o que acontecerá? E se parte de uma coleção digital for perdida, como por exemplo, através de um evento catastrófico?

Vê-se aí que questões de gerenciamento, acesso e preservação em longo prazo exigirão políticas e um esquema onde cópias permanentes redundantes são armazenadas em instituições designadas.

b) Tecnologias para digitalização, práticas de conversão e formatos de armazenamento de objetos de informação.

Quando um ambiente digital possui objetos de informação em formato original “tangível” ou analógico, tecnologias de digitalização deverão ser selecionadas cuidadosamente para que as informações possam oferecer qualidade, retratando fielmente os conteúdos dos objetos originais. Entretanto, grande parte destas tecnologias é emergente, forçando escolhas arriscadas que poderão não oferecer a qualidade desejada ou até mesmo por em risco a integridade física dos objetos originais (por exemplo, na conversão de elementos analógicos para digitais). Há poucas práticas ou padrões estabelecidos, além de escassez de ferramentas que proporcionem medições da qualidade de reprodução dos elementos digitalizados [Congress1].

Na criação de objetos de informação para ambiente digital, diretrizes também devem ser seguidas com relação a *softwares*, formatos e *layouts* dos objetos, visando tornar o gerenciamento da informação, a curto e longo prazo, mais fácil. Desta forma, destaca-se a importância nas práticas de captura e armazenamento do objeto de informação, para que não se corra o risco destes não poderem ser preservados, gerenciados e utilizados adequadamente

[FTFD1997]. Deve-se estar alerta com relação à escolha de suportes digitais para armazenamento (fitas, cartuchos, CDs, DVDs etc.), tanto quanto à expectativa de vida média destes, como também com relação à obsolescência de *hardware* e *software* que interagem com tais suportes.

- c) Ferramentas que proporcionem a geração automática de metadados úteis e facilitem a sua geração manual, baseando-se em vocabulário padrão de classificação.

Como visto anteriormente, metadados são importantíssimos em ambientes digitais, visto que são os elementos chave na busca e descoberta de informação, como também no gerenciamento de todo o seu ciclo de vida.

Metadados podem ser extraídos de forma manual ou automática. Muitas características de baixo nível podem ser extraídas de modo completamente automático, enquanto que características de alto nível normalmente necessitam da interação humana.

Observa-se, entretanto, que para a maioria dos objetos de informação, metadados continuam sendo criados manualmente e em momento posterior à sua criação. A criação manual de metadados é dispendiosa, além de passível de erros de digitação e emprego de vocabulário fora do padrão adotado [Hodge2000]. Uma vez que sistemas avançados de busca e descoberta de informação se baseiam em conteúdos de metadados, é necessário se buscar soluções que gerem automaticamente metadados e facilitem a sua geração manual, baseando-se em algum vocabulário padrão de classificação.

A necessidade de técnicas automatizadas para criação de metadados tem sido reconhecida por várias comunidades envolvidas no gerenciamento, produção, distribuição, uso e preservação de objetos de informação, e estudos ainda se encontram em fase inicial de pesquisa.

À medida que grupos de padrões e fabricantes de *softwares* procuram incorporar arquiteturas que facilitam a descrição de objetos de informação em seus produtos de processamento de textos, imagens e banco de dados como o XML (*eXtensible Mark-up Language*) e RDF (*Resource Description Framework*), a criação de metadados no momento da origem do objeto se tornará mais fácil.

- d) Identificadores e persistência de objetos de informação.

Este desafio também está relacionado com metadados, e trata do problema de nomear objetos de informação através de identificadores exclusivos. Tais identificadores são tão importantes em um ambiente digital, como o número de ISBN é em uma biblioteca tradicional, atendendo a propósitos como organização, recuperação de informação, formação de vínculos entre elementos, gerenciamento de direitos autorais etc.

Sistemas de identificação devem ser permanentes, devendo durar indefinidamente [Bullock1999]. Isto significa, entre outras coisas, que tais elementos não podem estar amarrados a uma localização específica, visto que esta pode ser alterada a qualquer momento. Este método é bem diferente do praticado atualmente na Internet, com o identificador URL (*Universal Resource Location*). O URL reúne, em uma única *string*, vários itens que poderiam estar separados: o protocolo necessário para acessar o objeto de informação (por exemplo, HTTP), o nome da máquina e o caminho para localização desse objeto, e o nome do arquivo que armazena o objeto (que pode ou não ser único). Dessa forma, métodos de identificação semelhantes ao URL podem trazer sérios problemas, visto que sempre que um arquivo é movido, o objeto de informação freqüentemente é perdido.

Exemplos de esquemas atuais de identificação única que não amarram localizações específicas ou processos são o PURL (URL persistente), o URN (*Universal Resource Name/Number*), e o Sistema DOI (Sistema de Identificadores de Objeto Digital). Nestes esquemas, identificadores permanecem válidos mesmo que objetos de informação sejam movidos de um local para outro, ou quando migrados de um meio de armazenamento a outro.

Assim, um dos fortes desafios a ser enfrentado por ambientes digitais, considerando que estes oferecerão interoperabilidade com outros ambientes afins, será administrar o sistema de nomes únicos de forma eficiente e efetiva.

e) Seleção de material para digitalização.

O valor da informação e os benefícios advindos dos acessos às versões digitais justificam os esforços na criação de um ambiente digital [Columbia2001]. Objetos de informação a serem digitalizados devem ter valor cultural e de pesquisa intrínseco suficiente para garantir o uso continuado pela comunidade atual e futura de usuários, por um período longo de tempo. Dessa forma, a atividade de seleção de material para digitalização é de

fundamental importância. Há diversas propostas para seleção de material para digitalização. Por exemplo:

- *Conversão retrospectiva de todas as coleções de uma instituição*, que apesar de ideal, é impraticável ou impossível tecnicamente, legalmente, e economicamente;
- *Digitalização de uma ou mais coleções especiais* (ou parte delas), quando estas são valiosas e de tamanho manejável;
- *Digitalização de parte de coleções*, selecionando apenas bons exemplos que realcem a riqueza da coleção;
- *Digitalização de materiais de alto uso*, aumentando a acessibilidade à informação de maior demanda;
- *Digitalização conforme solicitação*, onde se digitaliza e armazena objetos de informação à medida que estes são solicitados para uso. Este é um método de construção de coleções digitais ao acaso.

Estas propostas podem ser utilizadas separadamente ou em combinação, e esta escolha dependerá das políticas adotadas pela instituição para a geração de seu ambiente digital. Dentro de tais propostas há diversos critérios de seleção de material, os quais dependerão dos propósitos do ambiente digital. Por exemplo: seu potencial para uso em longo prazo; seu valor intelectual, cultural e físico; se ele proverá maior acesso do que já oferece como material original (por exemplo, materiais frágeis e raros); se restrições de direitos autorais permitirão conversão; se há exigência na legislação para disponibilizá-lo; se provê recuperação de custos; se há tecnologias disponíveis para digitalização e para visualização (viabilidade técnica) etc. [Harvard2004].

De acordo com o propósito do ambiente digital a ser construído, bem como com a natureza da instituição, critérios devem ser selecionados. Por exemplo, considerando uma instituição pública, e como tal inserida numa realidade de orçamentos limitados, seria interessante adotar critérios que favoreçam a recuperação de seus custos, bem como que promovam a captação de recursos, como por exemplo, critérios que promovam o “bem público” (isto é, aqueles que apóiam a educação formal e aprendizagem contínua, reforçam uma consciência nacional compartilhada e cidadania informada, e contribuem para o crescimento econômico e geração de trabalho).

f) Preservação das coleções digitais.

Outro desafio importante na construção de ambientes digitais é a preservação dos seus objetos, isto é, a conservação do conteúdo, da aparência e da percepção da informação digital, mantendo-a disponível em longo prazo.

Apesar de não haver nenhum consenso na definição de “preservação em longo prazo”, aqui esta definição pode ser imaginada como sendo uma preservação longa o suficiente para não haver preocupações com mudanças em tecnologia e na comunidade de usuários. Na preservação do material digital, o problema real está na obsolescência técnica, significando constantes migrações para novas soluções. Dependendo das tecnologias utilizadas e temas envolvidos na preservação, pode-se calcular um ciclo de migração de *hardware* e *software* de 2 a 10 anos, e de migração de mídias de armazenamento de 3 a 5 anos [Hodge2000].

A solução mais comum para o problema de mudança de mídia de armazenamento (fitas, discos, CDs etc.) é a migração para novos sistemas de armazenamento. Suportes de armazenamento têm um tempo de vida muito pequeno quando considerados em termos de obsolescência (em média de dois a cinco anos), exigindo sua migração para outras tecnologias mais atualizadas. Após esse período, objetos de informação que não sofreram o processo de migração correm o risco de serem perdidos, uma vez que poderá não mais existir *hardware* e/ou *software* para lê-los. Tal procedimento é oneroso, e sempre há preocupação sobre a perda de dados ou problemas com a qualidade quando uma transferência é realizada. Algoritmos de checagem são extremamente importantes quando esta proposta é utilizada.

Com relação a *hardware* e *software*, verifica-se que o mercado lança a cada 2 ou 3 anos novas tecnologias de bancos de dados, planilhas eletrônicas e processadores de textos, com *patches* e atualizações secundárias liberados de forma mais freqüente. Apesar de fabricantes de *software* geralmente proverem estratégias de migração ou compatibilidade para as novas gerações dos seus produtos, nem sempre isto acontece após uma ou duas gerações. Experiências mostram que migrações nem sempre garantem a cobertura de todos os tipos de dados armazenados em um sistema, e poderá haver perda de integridade da informação no resultado do processo.

Estratégias de migração de *hardware* e *software* são menos rigorosas que aquelas aplicadas em mídias de armazenamento. Para se evitar migrações freqüentes, as organizações têm procurado obter tecnologias comerciais populares. Como uma alternativa à migração, atualmente está começando a ser considerada a emulação do comportamento do *hardware/software* encapsulada com o objeto de informação. Por exemplo, um documento MS Word 2000 poderia ter em seus metadados informações que indicassem como reconstruí-lo em nível de engenharia – *bits* e *bytes*.

Em [HR1998; Rothenb1999] encontra-se uma alternativa para encapsular o *software* com todas as instâncias de tipos de dados, criando um registro de emulação que exclusivamente identifica os ambientes de *hardware* e de *software* e provê informação de como recriar tais ambientes para preservar o uso do objeto de informação. Até o momento, não há nenhum sistema que proveja documentação e informação de emulação necessária para esta proposta ser operável. Mais importante, não existe nenhuma política que exija dos fabricantes a disponibilização de informação de emulação. Dessa forma, a melhor prática para os próximos anos continuará sendo a migração a novas plataformas de *hardware* e de *software*; a emulação começará a ser usada se (ou quando) as indústrias de *hardware* e de *software* começarem a endossar esse processo.

A preservação de formatos disponíveis (tamanhos de bloco, tamanhos de fitas, mecanismos de transmissão etc.) está relacionada à continuidade de acesso ao conteúdo digital. Enquanto arquivos podem ser movidos de um meio de armazenamento físico para outro, o que acontece quando o formato da informação (por exemplo, PDF) fica obsoleto? Uma solução é fazer migração de dados (traduzir os dados de um formato para outro). Entretanto, há também aqui dificuldades: a migração de dados é muito cara, não há ainda nenhum padrão para esse processo, e distorção ou perda de informação é inevitavelmente introduzida.

A.3 Desafios na Interoperabilidade de Ambientes Digitais:

- a) Estabelecer protocolos, padrões e práticas que viabilizem a interoperabilidade, compartilhamento, integração e reuso de informação entre ambientes digitais remotos de entidades afins.

É desejável que ambientes digitais remotos pertencentes a entidades afins possam se interoperar, gerando uma unidade virtual [Congress1]. Para tanto, é necessário que se busquem soluções para várias questões importantes, como por exemplo:

- Eleger protocolos e práticas, e definir padrões para os objetos de informação digital, de forma a alcançar um equilíbrio entre viabilidade de implementação distribuída, coerência de acesso e preservação coordenada;
- Definir padrões compreensíveis por pessoas e máquinas para descrição de recursos na *Web*, que proporcionem automação, compartilhamento, integração e reuso de informação em vários níveis e a várias aplicações;
- Buscar solução para, em ambiente digital distribuído, proteger os direitos associados ao conteúdo (direitos de privacidade e condições impostos por doadores, bem como direitos autorais), enquanto ainda provendo acesso o mais amplo possível.

Este é um desafio que a proposta da Web Semântica, com todas as suas tecnologias associadas, vêm contribuir fortemente. Segundo a W3C [W3C-SW], a Web Semântica tem o objetivo de criar um meio universal para a troca de dados, apropriado tanto para pessoas como para máquinas, dentro dos mais variados propósitos, para que a *Web* alcance seu potencial completo.

A viabilidade dos dados se tornarem compreensíveis por máquinas na *Web* está se tornando rapidamente uma prioridade para as organizações, indivíduos e comunidades que desejam que seus dados sejam explorados pelas máquinas não só com o objetivo de apresentação, mas também para automação, integração e reuso de dados entre aplicações.

Entre as tecnologias que fazem parte da Web Semântica encontra-se o *Resource Description Framework* (RDF), que apóia descrição de recursos para descoberta de informação, bem como para gerenciamento de direitos autorais, preferências de privacidade, avaliação de conteúdos e classificação. RDF é visto como o esquema para produzir uma *Web* de confiança, onde o conteúdo de cada objeto individualmente acessível é bem descrito em um formato que é extensível, contudo universalmente entendido. RDF pode capacitar máquinas de procura e outras ferramentas para descoberta de recurso, na troca e

compartilhamento de meta dados, seguindo os princípios do W3C de interoperabilidade, evolução, e descentralização.

b) Acompanhar as mudanças em tecnologias de informação e comunicação (TIC)

A disponibilidade crescente de padrões em TIC tem o potencial de reduzir a interoperabilidade e interconectividade de usuários e ambientes, devido à dificuldade em acompanhar o ritmo acelerado das mudanças. As diversas opções em padrões e protocolos de rede, inovações tecnológicas e novos tipos de mídia e formatos exigirão trocas periódicas (às vezes radicais) na forma de gerenciar e prover acesso às informações.

A.4 Desafios em Direitos Autorais:

a) Provisão ou restrição de acesso de forma adequada, variação do direito de acesso de acordo com os direitos de propriedade intelectual do material, e variação dos níveis de segurança.

Um dos mais importantes desafios na construção de ambientes digitais tem sido a questão da proteção de direitos autorais para objetos de informação, o qual inclui provisão ou restrição de acesso de forma adequada, variação do direito de acesso de acordo com os direitos de propriedade intelectual do material, e variação dos níveis de segurança. Se por um lado tais ambientes oferecem grandes oportunidades em disseminar informação, por outro eles aumentam as dificuldades envolvidas no gerenciamento e proteção dos direitos autorais [Cleve1998].

Profissionais envolvidos em digitalização e armazenamento digital preocupam-se com questões como [Hodge2000]: Quais os direitos de um arquivo? Quais os direitos dos vários grupos de usuário? Quais os direitos que um proprietário deve reter? Como mecanismos de acesso podem interagir com metadados de arquivo para assegurar que direitos de propriedade serão administrados corretamente?

O problema para a maioria dos mantenedores de informação é que em geral eles simplesmente são guardiões das informações, não possuindo os direitos autorais sobre o material que eles mantêm. É bem improvável que tais entidades consigam digitalizar

livremente e prover acesso a todos os objetos de informação com direitos autorais das suas coleções.

Independente do formato que um trabalho se encontre (digital ou não), a legislação atual prevê a mesma proteção de direito autoral, isto é, as mesmas regras se aplicam tanto dentro como fora de um ambiente digital. Entretanto, o conceito atual baseado em papel para tal proteção não funciona em tais ambientes, visto que não há como controlar o número de cópias dos objetos.

Segurança e controle de versões também são complicados em sistemas digitais, visto que estas são menos estáticas, facilmente copiadas, e remotamente acessíveis por múltiplos usuários simultaneamente. Várias questões são levantadas com relação à privacidade e roubo de informação, particularmente quando se é comum armazenar todos os *sites* que estão associados entre si, formando uma longa cadeia. Semelhante preocupação há entre arquivistas de imagens, uma vez que estas podem ser manipuladas sem a falsificação ser descoberta.

Assim, são vários os desafios em direitos autorais a serem enfrentados, e dentre eles pode-se citar [FTFD1997]:

- Identificação da abrangência das questões de direitos autorais nas informações a serem disponibilizadas.
 - O material a ser disponibilizado está protegido por direitos autorais?
 - Se sim, quais os direitos envolvidos na defesa da utilização pretendida?
 - Os direitos apropriados ou licenças na utilização do material já foram adquiridos?
 - Se não, quem possui os direitos?
- Mecanismos de liberação de direitos.
 - Quando os direitos autorais dos trabalhos não pertencem à Instituição, como poderão ser localizados seus portadores de propriedade intelectual?
 - Quando os direitos autorais pertencem à Instituição, como serão avaliados os materiais que necessitam de liberação?
- Gerenciamento de direitos.

- Como prover tratamento das transações com usuários, gerenciando número de cópias, fazendo rastreamento do uso etc.?
- O que pode ser feito para assegurar que usuários estão sendo informados sobre trabalhos protegidos por direitos autorais (provisão da situação de *copyright* de cada objeto de informação, provisão das restrições ou das taxas associadas para uso)?
- Quais tecnologias (por exemplo, criptografia, marca d'água, assinatura digital, identificação e autenticação de usuários) ajudariam a melhorar a proteção e facilitariam o gerenciamento de direitos autorais?
- Como podem os direitos autorais de um trabalho serem protegidos utilizando os meios legais existentes (por exemplo, através de lei de propriedade intelectual, através de contrato etc.)?
- Necessidade de conscientização sobre direitos autorais.
 - Como pode ser aumentada a conscientização sobre direitos autorais na sociedade brasileira?

Metadados permitem registrar informação de direitos de propriedade intelectual e reprodução, atribuídos aos objetos de informação e suas múltiplas versões. Metadados também documentam outras exigências legais ou de doadores, que tenham sido impostas aos objetos de informação, para, por exemplo, atender a questões de privacidade ou de interesse de proprietário. Particularmente em casos onde assuntos de proteção estão em risco, é importante se ter metadados para gerenciar criptografia, marcas d'água, assinaturas digitais etc., que podem sobreviver apesar de mudanças no formato de armazenamento e nas mídias onde os objetos de informação são armazenados.

A.5 Desafios na Formação e Sustentação de Recursos Humanos:

- a) Formar e sustentar recursos humanos competentes e comprometidos com o ambiente digital.

A geração de um ambiente digital exigirá o envolvimento de grupos de profissionais inter-disciplinares que necessitarão de capacitação específica na área que atuarão, e

compromisso no cumprimento de suas funções. Este é mais um dos grandes desafios a ser enfrentado, visto que tais profissionais serão os elementos que determinarão o nível de qualidade do ambiente digital [Congress2]. Dessa forma, é de extrema importância que se invista na formação e manutenção de tais recursos humanos.

A.6 Desafios na Captação de Recursos Financeiros:

- a) Captar recursos financeiros para a criação, ampliação e manutenção de toda a infraestrutura geral exigida para a criação do ambiente digital semântico, que compreende:
- Mobilização e motivação de recursos humanos em todas as suas áreas de atuação;
 - Infra-estrutura tecnológica para implementação e suporte de todos os seus serviços;
 - Construção de espaços físicos com infra-estrutura apropriada para acomodação adequada de toda a infra-estrutura tecnológica, grupos de trabalho, e ambientes para oferecer acessos;
 - Sistemas efetivos de segurança física que assegurem acesso adequado e ao mesmo tempo provejam proteção para o pessoal, instalações, dados e coleções.
- b) Levantamento de modelos econômicos para sustentar o ambiente digital.

A produção e manutenção de ambientes digitais é muito cara. Seu custo decorre da produção, organização, armazenamento, provisão contínua de acesso, distribuição, e preservação da informação digital. A sustentação de tais ambientes necessita de melhores estimativas de custos e tendências para produzir e manter um corpo de informação digital [Congress1]. Dessa forma, a Instituição deverá buscar ferramentas que lhe permitam obter um conhecimento melhor dos custos contínuos de criação e manutenção de ambientes digitais, verificando também se há tecnologia disponível que ofereça melhores medições de vantagens e economias. Neste momento será importante o intercâmbio com instituições afins que possam contribuir com experiências e soluções.

c) Desenvolver estratégias de captação de recursos.

Além das estimativas de custos, a Instituição deverá buscar meios de conseguir financiamentos. Deverá contar também com esquemas de parcerias e estratégias de auto-sustentação, onde poderá oferecer, além de serviços gratuitos, classes diferenciadas de serviços com privilégios, como por exemplo: conexão de rede com qualidade diferenciada, área privada personalizada para acesso, acesso a coleções especiais, acesso a diferentes versões de informação com qualidade, formato e nível de detalhamento diferenciado, uso de ferramentas especiais de aprimoramento de buscas e recuperação de informação etc. Tais serviços diferenciados deverão estar disponíveis via contrato formal, ou por meio de métodos de pagamento *on-line* para a cobrança automatizada de acesso aos objetos de informação que prevejam custos. Outras formas de arrecadação de recursos financeiros poderiam ser criadas como, por exemplo, oferecer serviços de geração e/ou hospedagem de coleções digitais para outras instituições.

Anexo A

Diagramas Comportamentais do Modelo SIGAMeta

Este Anexo apresenta o Diagrama de Caso de Uso e o Diagrama de Seqüência (Figura A.2) do Modelo proposto nesta Tese. Estes foram escritos na Linguagem UML (*Unified Modelling Language*), utilizando-se a ferramenta de modelagem de sistemas *Poseidon for UML – Community Edition* (<http://www.mondo.com.br/modelagem/Poseidon.shtml>). A Linguagem UML é uma linguagem de diagramação ou notação para especificar, visualizar e documentar modelos de sistemas de *software* orientados a objeto. A UML é controlada pelo *Object Management Group* - OMG (<http://www.uml.org/>), e é um padrão da indústria para descrever graficamente *software*.

A.1 Diagrama de Caso de Uso

O Diagrama de Caso de Uso (Figura A.1) representa os pontos de interação entre o usuário e o sistema. Ele mostra atores (pessoas ou outros usuários do sistema), casos de uso (os cenários onde eles usam o sistema), e seus relacionamentos.

Figura A.1 – Diagrama de Caso de Uso

A.2 Diagrama de Sequência

O Diagrama de sequência é um tipo de diagrama usado em UML que representa a sequência de processos (mais especificamente, de mensagens passadas entre objetos), num programa de computador. Ele descreve a maneira como os grupos de objetos colaboram em algum comportamento ao longo do tempo, e registra o comportamento de um único caso de uso. Em Diagramas de Sequência, objetos são representados através de linhas verticais tracejadas, com o nome do Objeto no topo. O eixo do tempo é também vertical, aumentando para baixo, de modo que as mensagens são enviadas de um Objeto para outro na forma de setas, com a operação e os nomes dos parâmetros.

As figuras A.2 e A.3 abaixo exibem os diagramas de sequência respectivamente para as operações *processar categoria* (Fase 1 - Treinamento) e *processar documento* (Fase 2 – Geração de Metadados).

Figura A.2 – Diagrama de Sequência da Operação *Processar Categoria*

Figura A.3 – Diagrama de Seqüência da Operação *Processar Documento*

Anexo B

Informações e Resultados Complementares

Este Anexo apresenta algumas listas, tabelas, figuras e resultados complementares obtidos nos experimentos do Modelo SIGAMeta.

B.1 Lista de *Stopwords*

A Lista de *stopwords* foi composta por 2.679 termos gerados manualmente. Os termos começados com a letra “a” encontram-se abaixo.

Tabela B.1 – Lista de *Stopwords*

a	à	abaixo	abordagem	abordagens	abordar
abril	absoluta	absolutamente	absoluto	acaba	acabado
acabai	acabais	acabam	acabamos	acabámos	acabando
acabar	acabara	acabará	acabaram	acabáramos	acabarão
acabaras	acabarás	acabardes	acabarei	acabareis	acabáreis
acabarem	acabaremos	acabares	acabaria	acabariam	acabáramos
acabarias	acabáreis	acabarmos	acabas	acabasse	acabásseis
acabassem	acabássemos	acabasses	acabaste	acabastes	acabava
acabavam	acabávamos	acabavas	acabáveis	acabe	acabei
acabeis	acabem	acabemos	acabes	acabo	acabou
acaso	aceita	aceitado	aceitai	aceitais	aceitam
aceitámos	aceitando	aceitar	aceitara	aceitará	aceitaram
aceitáramos	aceitarão	aceitaras	aceitarás	aceitardes	aceitarei
aceitareis	aceitáreis	aceitarem	aceitaremos	aceitares	aceitaria
aceitariam	aceitariamos	aceitarias	aceitáreis	aceitarmos	aceitas
aceitasse	aceitásseis	aceitassem	aceitássemos	aceitasses	aceitaste
aceitastes	aceitava	aceitavam	aceitávamos	aceitavas	aceitáveis
aceite	aceitei	aceiteis	aceitem	aceitemos	aceites
aceito	aceitou	acerca	acesso	acima	acinte
acompanhado	acompanhar	acompanharam	acompanhou	aconteça	aconteçam
acontece	acontecem	acontecendo	acontecer	acontecera	acontecerá
aconteceram	acontecerão	acontecerem	aconteceria	aconteceriam	acontecesse
acontecessem	aconteceu	acontecia	aconteciam	acontecido	acontecimento
adentro	adequação	adiante	adição	adicionado	adicionar
adicionava	adota	adotada	adotadas	adotado	adotados
adotamos	adotar	adotara	adotou	adrede	afetam
afetar	afetaram	afetou	afinal	afirma	afirmação
afirmada	afirmando	afirmar	afirmava	afirmavam	afirmou

afora	agora	agosto	aí	ainda	al
além	alemães	alemanha	alemão	algo	alguém
algum	alguma	algumas	alguns	algures	alhures
ali	aliada	aliadas	aliado	aliados	aliás
alta	alteração	alterações	alterar	alternativa	alternativo
alto	amanhã	ambas	ambos	amiúde	amplamente
amplas	ampliação	ampliando	ampliar	ampliaram	ampliou
ano	anos	ante	anteontem	anterior	anteriores
anteriormente	antes	ao	aonde	aos	apenas
apesar	aponta	apontados	apontaram	apontava	apontavam
após	apresenta	apresentação	apresentações	apresentado	apresentai
apresentais	apresentam	apresentamos	apresentámos	apresentando	apresentar
apresentara	apresentará	apresentaram	apresentáramos	apresentarão	apresentaras
apresentarás	apresentardes	apresentarei	apresentáreis	apresentarem	apresentaremos
apresentares	apresentaria	apresentariam	apresentaríamos	apresentarias	apresentarieis
apresentarmos	apresentas	apresentasse	apresentásseis	apresentassem	apresentássemos
apresentasses	apresentaste	apresentastes	apresentava	apresentavam	apresentávamos
apresentavas	apresentáveis	apresente	apresentei	apresenteis	apresentem
apresentemos	apresentes	apresento	apresentou	aquela	àquela
aquelas	àquelas	aquele	àquele	aqueles	àqueles
aquém	aqui	aquilo	àquilo	art	as
às	aspecto	aspectos	assaz	assim	assinala
assinalado	assinalai	assinalais	assinalam	assinalamos	assinalámos
assinalando	assinalar	assinalara	assinalará	assinalaram	assinaláramos
assinalarão	assinalaras	assinalarás	assinalardes	assinalarei	assinalareis
assinaláreis	assinalarem	assinalaremos	assinalares	assinalaria	assinalariam
assinalaríamos	assinalarias	assinalaríeis	assinalarmos	assinalas	assinalasse
assinalásseis	assinalassem	assinalássemos	assinalasses	assinalaste	assinalastes
assinalava	assinalavam	assinalávamos	assinalavas	assinaláveis	assinale
assinalei	assinaleis	assinalem	assinalemos	assinales	assinalo
assinalou	assuma	assumais	assumam	assumamos	assumas
assume	assumem	assumes	assumi	assumia	assumiam
assumíamos	assumias	assumido	assumíeis	assumimos	assumindo
assumir	assumira	assumirá	assumiram	assumíramos	assumirão
assumiras	assumirás	assumirdes	assumirei	assumireis	assumíreis
assumirem	assumiremos	assumires	assumiria	assumiriam	assumiríamos
assumirias	assumiríeis	assumirmos	assumis	assumisse	assumísseis
assumissem	assumíssemos	assumisses	assumiste	assumistes	assumiu
assumo	assunto	assuntos	até	atenção	atendendo
atender	atenderia	atendeu	atendida	atendidas	atendido
atendimento	atingida	atingidas	atingido	atingidos	atingir
atingiu	ator	atores	atrair	atrás	através
atribuição	atribuições	atribuída	atribuídas	atribuído	atribuídos
atribuir	atualidade	autônoma	autônomas	autonomia	autônomo
autônomos	autor	autora	autores	autoria	autoria
Avançadas	avançar	avançavam	avanço		

B.2 Vetores de Categorias

O vetor de categoria CPOLI-RSLP foi composto por 390 *stemmings*, e o CPOLI-Porter foi composto por 392. Abaixo se encontra uma lista com as cinquenta primeiras palavras desses vetores, ordenadas decrescentemente por peso.

Tabela B.2 – Vetor de Categoria CPOLI

	<i>Stemming RSLP</i>	<i>Peso</i>	<i>Stemming Porter</i>	<i>Peso</i>
1	polit	0,2246204026133631	polít	0,23430952215963088
2	govern	0,18846437461276475	govern	0,19430292653505268
3	brasil	0,10318175371796491	estad	0,11779853277711029
4	presid	0,10167994098086017	trabalh	0,09040798609687557
5	est	0,08966530078194465	país	0,08744261930652121
6	trabalh	0,08750279572420754	milit	0,08722392706789388
7	país	0,08620384673429206	varg	0,08664743322929161
8	milit	0,08560102588485818	president	0,0739048462728541
9	varg	0,08127874508792983	part	0,06844199801759751
10	part	0,06834446396571828	grand	0,05544840037925749
11	nacion	0,06621535606767742	nacional	0,04990331258985115
12	grand	0,05201547477490447	brasileir	0,04957729411788139
13	sindic	0,05044209389545527	jang	0,04670517828887459
14	candidat	0,0457658094556848	moviment	0,04651934386700255
15	jang	0,04572595180610832	apoi	0,041158320709653744
16	mov	0,04343793052710937	brasil	0,03874867799720277
17	apoi	0,038669863948112566	constituiçã	0,03823464008399183
18	goulart	0,036926011220883126	goulart	0,037981584646191814
19	histor	0,03565200894035182	sindical	0,03586416754835196
20	constitu	0,035634427969113674	histór	0,03545271258896293
21	ministr	0,03168219187735337	ministr	0,032143519855546386
22	congress	0,031119156724841814	congress	0,0318578016872072
23	soc	0,029101599845860613	social	0,030949816721932966
24	ptb	0,028338746513562364	ptb	0,030734617705241763
25	econom	0,02816454143594183	econôm	0,029312202856907527
26	sindicat	0,026434132649147932	sindicat	0,027166560083318716
27	grev	0,0261147390246831	comun	0,02596312876332727
28	udn	0,02401812201450678	udn	0,0256234250552548
29	psd	0,02352848531460857	psd	0,02462410177765314
30	getuli	0,022836787095492982	grev	0,02440906314244776
31	opos	0,021901586811412813	democrát	0,022254486071905565
32	comun	0,021332087756155495	forc	0,022205182196305212
33	jk	0,020003838521033202	jk	0,021113800631057014

34	regim	0,019952087174498024	regim	0,020558475216511687
35	democr	0,019910715931895813	desenvolv	0,019808058091498278
36	intelect	0,01968485734127869	organiz	0,018591007866786843
37	desenvolv	0,01966687012310726	jâni	0,01836411150431773
38	ele	0,019379044097861998	candidatur	0,01823544920235145
39	forç	0,01927979780757911	president republ	0,016790929948541814
40	public	0,018154774468257025	vot	0,015953236656531835
41	organiz	0,017816690006815353	popul	0,015810899634971384
42	cult	0,017761480077068913	frent	0,01439954677087763
43	jani	0,0174807721675103	eleitoral	0,014123493258937683
44	feder	0,017153629699192076	reform	0,014054732939173769
45	golp	0,016316908209073555	histór polít	0,013985758991520906
46	paul	0,016063793187151634	grup	0,013888066613566752
47	presid republic	0,015688707102948122	oposiçã	0,013308301888104744
48	vot	0,015569298585247453	poss	0,01324215546837392
49	eleitor	0,01430190367340858	revoluçã	0,01278886168901316
50	grup	0,013856980170694732	eleiçõ	0,012571855352344006

O vetor de categoria ECONO-RSLP foi composto por 1.008 *stemmings*, e o ECONO-Porter foi composto por 1.027. Abaixo se encontra uma lista com as cinquenta primeiras palavras desses vetores, ordenadas decrescentemente por peso.

Tabela B.3 – Vetor de Categoria ECONO

	<i>Stemming RSLP</i>	<i>Peso</i>	<i>Stemming Porter</i>	<i>Peso</i>
1	empres	0,08522062510764664	empres	0,08522062510764664
2	econom	0,08483969890555615	econom	0,08483969890555615
3	capital	0,07464943377571844	capital	0,07464943377571844
4	país	0,06949088872903733	país	0,06949088872903733
5	produt	0,06827276975613185	produt	0,06827276975613185
6	desenvolv	0,06281710089743169	desenvolv	0,06281710089743169
7	histór	0,0562397385283632	histór	0,0562397385283632
8	grand	0,05505232954123722	grand	0,05505232954123722
9	merc	0,052627599340190886	merc	0,052627599340190886
10	mudanc	0,051046178963364365	mudanc	0,051046178963364365
11	process	0,04845522312695628	process	0,04845522312695628
12	anális	0,04685132729000499	anális	0,04685132729000499

13	produçã	0,04647231585010304	produçã	0,04647231585010304
14	valor	0,04434514973289479	valor	0,04434514973289479
15	europ	0,04412552290982626	europ	0,04412552290982626
16	sistem	0,044051506397618824	sistem	0,044051506397618824
17	trabalh	0,0434025130241411	trabalh	0,0434025130241411
18	cas	0,04307488103907569	cas	0,04307488103907569
19	sociedad	0,042560954982229025	sociedad	0,042560954982229025
20	aument	0,0417630184389919	aument	0,0417630184389919
21	reassent	0,04080187858007666	reassent	0,04080187858007666
22	chesf	0,04068989894319379	chesf	0,04068989894319379
23	econôm	0,03860091603633344	econôm	0,03860091603633344
24	comun	0,03698294463597791	comun	0,03698294463597791
25	feudal	0,03663991659523775	feudal	0,03663991659523775
26	estad	0,036445265519194164	estad	0,036445265519194164
27	model	0,03364677153534236	model	0,03364677153534236
28	banc central	0,0326874083411851	banc central	0,0326874083411851
29	maior	0,032670300390008566	maior	0,032670300390008566
30	total	0,0313572137708601	total	0,0313572137708601
31	moed	0,029538436515322085	moed	0,029538436515322085
32	polít	0,02932779464445875	polít	0,02932779464445875
33	elev	0,028092580830942	elev	0,028092580830942
34	particip	0,027180463974614197	particip	0,027180463974614197
35	oper	0,026609960982933122	oper	0,026609960982933122
36	famil	0,026323736808157933	famil	0,026323736808157933
37	cust	0,02591974599929586	cust	0,02591974599929586
38	local	0,025722437926234264	local	0,025722437926234264
39	empreg	0,025707228834768104	empreg	0,025707228834768104
40	pequen	0,025699209042237596	pequen	0,025699209042237596
41	invest	0,025440569422943957	invest	0,025440569422943957
42	projet	0,025428453309778346	projet	0,025428453309778346
43	inform	0,025044397299020284	inform	0,025044397299020284
44	cris	0,024444141804127923	cris	0,024444141804127923
45	áre	0,024044454438765378	áre	0,024044454438765378
46	prec	0,023972265017894966	prec	0,023972265017894966
47	banc	0,023954495827406254	banc	0,023954495827406254
48	reduz	0,02327934357134649	reduz	0,02327934357134649
49	firm	0,022922212350849555	firm	0,022922212350849555
50	plan	0,022683395990820646	plan	0,022683395990820646

O vetor de categoria ESTPOP-RSLP foi composto por 824 *stemmings*, e o ESTPOP-Porter foi composto por 834. Abaixo se encontra uma lista com as cinquenta primeiras palavras desses vetores, ordenadas decrescentemente por peso.

Tabela B.4 – Vetor de Categoria ESTPOP

	<i>Stemming RSLP</i>	<i>Peso</i>	<i>Stemming Porter</i>	<i>Peso</i>
1	mulh	0,16994960013369317	mulh	0,17290625558694153
2	trabalh	0,16178526418252928	trabalh	0,16238092618928088
3	analís	0,07465776273336182	famíl	0,10577139291855174
4	migr	0,07242701821959494	áre	0,07057359333156131
5	popul	0,06911143741470743	popul	0,06983282610952657
6	famil	0,06790235104345187	populacion abep	0,06223385179466011
7	econom	0,06733037028997128	peesso	0,058993471241080786
8	cresc	0,06411662543560465	regiã	0,05772828967494396
9	populac abep	0,062349315892529435	grand	0,05723629993965589
10	are	0,05952403663373999	estad	0,054556111479935415
11	pess	0,05787469620622681	migraçã	0,052250156834444866
12	grand	0,057726044723318534	homens	0,05132185881834357
13	hom	0,05659273771372163	migrant	0,05030071164010636
14	cond	0,05195953524403203	econôm	0,049788591644067896
15	pais	0,049551015825662285	país	0,049502724993013295
16	inform	0,04851259526239688	saúd	0,0487398118356589
17	saud	0,04783501272713555	diferenc	0,046305033782815884
18	sex	0,046886758265750096	municípi	0,04593695316641544
19	indic	0,04664214982747854	process	0,045087637574533726
20	municipi	0,04578080902156623	ativ	0,04343404071284773
21	brasil	0,0446733728609843	aument	0,04149254953292687
22	urban	0,043760691185566664	númer	0,04093475166441938
23	numer	0,04228143625987434	mortal	0,03952377546299701
24	process	0,04188315507893447	imigr	0,03843089445066669
25	aument	0,04123582385509643	anális	0,038334150585802064
26	reg	0,03907278108284783	total	0,037662815524628156
27	imigr	0,03830739128338618	inform	0,03765366342079907
28	paul	0,03827278360023799	indic	0,037505982115450656
29	famili	0,03784204931258295	famili	0,03726691039597223
30	mortal	0,03759659440967326	maior	0,03665399707648766
31	rend	0,03674303945096019	social	0,03626423251084838
32	total	0,03658060218810852	brasileir	0,03341531679046973

33	resid	0,036281219223881646	feminin	0,0331883992486974
34	soc	0,033754546892905	idos	0,03297517181677778
35	ativ	0,0333907843416168	polít	0,03222734014079
36	local	0,03314341880500077	ciudad	0,031900563953996396
37	cas	0,032991783642248904	desenvolv	0,03139882913922553
38	feminin	0,032952421791646584	merc trabalh	0,030993240426632037
39	ciudad	0,0324468075818586	mudanc	0,03083451775736395
40	idos	0,03227265105958383	condiçõ	0,03076163730287511
41	mudanç	0,03141147913998897	regiõ	0,02975207588103423
42	desenvolv	0,031151965351373236	filh	0,029700704663647454
43	merc trabalh	0,030187813960519217	cas	0,02967233482058564
44	tend	0,029306678022931586	cresciment	0,029212610362516527
45	filh	0,02917304475082814	jovens	0,02851076697266454
46	jov	0,02825706634612182	paul	0,028364497630754538
47	elev	0,028087945640980643	elev	0,027734264964593382
48	grup	0,027103753381433653	grup	0,02717027653418416
49	constitu	0,02694045646024007	rend	0,02609158689300945
50	escolar	0,026351781678602738	particip	0,025751272853741677

O vetor de categoria ANTROP-RSLP foi composto por 489 *stemmings*, e o ANTROP-Porter foi composto por 500. Abaixo se encontra uma lista com as cinquenta primeiras palavras desses vetores, ordenadas decrescentemente por peso.

Tabela B.5 – Vetor de Categoria ANTROP

	<i>Stemming RSLP</i>	<i>Peso</i>	<i>Stemming Porter</i>	<i>Peso</i>
1	danç	0,16412275856894537	danc	0,147389877276928
2	cult	0,0648589490668113	fest	0,06133146550688447
3	fest	0,06356523888107604	pov	0,05422773405965506
4	folclor	0,05893565613347529	cant	0,0484321454588084
5	pov	0,05323351718432739	folclor	0,04829955467449305
6	cant	0,04939378090182284	grup	0,041221801663064
7	grup	0,04069002554616115	cultur	0,035931145171766626
8	brasil	0,035781168573381344	grand	0,03225934252908528
9	cas	0,03487351332870972	sant	0,03121610844913084
10	sant	0,03395321013488269	pezzo	0,030716777214818775
11	grand	0,03385988083036013	branc	0,029735304300833012
12	branc	0,0280856067410261	brasil	0,028009088769158856

13	orix	0,02575082600779043	mulh	0,025296217571302113
14	lingu	0,02510914621832641	execut	0,024718973665922912
15	mulh	0,024850337493166348	popul	0,024468911170207348
16	pess	0,024654805008386346	noit	0,023605146911281433
17	popul	0,02421057997118195	bat	0,022582393164658158
18	execut	0,023668363799124157	menin	0,022079459014733006
19	noit	0,023361940042624325	cas	0,02186703840961011
20	mat	0,022256241158307592	par	0,02061472471856561
21	coreograf	0,02219946374721708	fandang	0,020588590984622002
22	music	0,02136561324882628	mat	0,019828971897143582
23	menin	0,020754913702338774	língu	0,01937891392801849
24	fandang	0,019542617885423722	trabalh	0,019029368066147162
25	candomble	0,019325591887687617	mit	0,018760862125741972
26	par	0,019221312439096746	brasileir	0,018605280640090392
27	trabalh	0,018774183394178714	mei	0,01833047066306057
28	negr	0,01854665949801528	human	0,01821962941432179
29	simbol	0,0183799420853555	artesanat	0,018117964865139235
30	mit	0,017795850598978546	dançador	0,01785906169918077
31	jog	0,01738439399397155	ciudad	0,017384773795989978
32	viv	0,01737119886704276	orixás	0,01702871829599615
33	artesanat	0,01735976350534172	viv	0,016784817374222872
34	viol	0,017007120340964674	carnaval	0,016654510119546117
35	ciudad	0,016768415498515377	palavr	0,016640219910357402
36	mei	0,01673234557914418	jog	0,016493751224684217
37	human	0,016662662030372363	noiv	0,01582954150110461
38	carnav	0,0165006625279106	sant antôni	0,015606245200175586
39	palavr	0,016084546005897087	joã	0,015068900181853966
40	sant antoni	0,016030252673706582	cong	0,014675109216204352
41	bat	0,015265469863384775	sociedad	0,013647557692766777
42	noiv	0,014937039358790153	comunic	0,013568346584732707
43	filh	0,014897455902851508	fogueir	0,013382456001917389
44	joao	0,01457421225843795	desfil	0,013259436194123439
45	hom	0,014316664278729735	músic	0,013127538510247913
46	local	0,013796616030938999	sab	0,012876711480039315
47	espirit	0,01364719708140284	cor	0,01257626259248708
48	sociedad	0,013297763905806224	águ	0,012575201073086708
49	soc	0,01317066465714371	fol	0,012408830810744708
50	rit	0,013102797643956695	negr	0,012299445660353008

B.3 Palavras e Stemming RSLP e Porter

O *corpus* da categoria Antropologia Cultural produziu um total de 11.692 palavras distintas. Abaixo se encontra uma lista com algumas palavras e seus respectivos *stemmings* para RSLP e Porter, em ordem alfabética.

Tabela B.6 – Palavras e *Stemmings* da Categoria ANTROP

<i>Palavra</i>	<i>Stemming RSLP</i>	<i>Stemming Porter</i>
abacaxi	abacax	abacax
abadá	abad	abad
abafo	abaf	abaf
abala	abal	abal
abaló	abalo	abal
abaluaê	abalua	abalua
abanando	aban	aban
abandona	abandon	abandon
abandonada	abandon	abandon
abandonado	abandon	abandon
abandonados	abandon	abandon
abandonando	abandon	abandon
abandonou	abandon	abandon
abanheenga	abanheeng	abanheeng
abanos	aban	aban
abarrotada	abarrot	abarrot
abater	abat	abat
abebé	abebe	abeb
abebê	abeb	abeb
abelhas	abelh	abelh
abençoa	abenç	abenço
abençoadado	abenço	abenço
abençoados	abenço	abenço
abençoaram	abenço	abenço
aberm	aberm	aberm
aberta	abert	abert
abertas	abert	abert
aberto	abert	abert
abertos	abert	abert
abertura	abert	abertur
abestado	abest	abest
abismo	abism	abism
abocanh	abocanh	abocanh
abolição	abol	aboliçã

abolicionistas	abol	abolicion
abolida	abol	abol
abolir	abol	abol
abordadas	abord	abord
abordado	abord	abord
abordará	abord	abord
aborígenes	aborigin	aborigin
aborrecem	aborrec	aborrec
abra	abr	abra
abraçar	abraç	abrac
abraçaram	abraç	abrac
abraçou	abraç	abrac
belas	bel	bel
belém	belem	belém
belezas	bel	belez
beligerância	beliger	beliger
belíssima	belissim	belíssim
belíssimas	belissim	belíssim
belo	bel	bel
belos	bel	bel
bênção	benç	bênçã
benedito	benedit	benedit
beneficiava	benefic	benefic
beneficiavam	benefici	benefic
benefício	benefici	benefici
benefícios	benefici	benefici
benesses	ben	ben
benevolência	benevol	benevolent
benfeitor	benfeitor	benfeitor
bengala	bengal	bengal
benjamin	benjamin	benjamin
bens	bem	bens
benta	bent	bent
bentinho	bent	bentinh
bento	bent	bent
benzedô	benzedo	benzedô
benzedores	benz	benzedor
benzer	benz	benz
beozzo	beozz	beozz
bepkororoti	bepkororot	bepkororot
berço	berç	berc

berços	berç	berc
berimbau	berimbau	berimbau
berlim	berlim	berlim
bernard	bernard	bernard
berra	berr	berr
berrantes	berr	berrant
berros	berr	berr
berta	bert	bert

B.4 Textos Originais

Abaixo se encontram alguns textos originais utilizados no experimento. Suas tabelas, autores e referências bibliográficas foram retirados para efeito de redução de espaço:

- **Traços gerais do desenvolvimento recente do Nordeste - Categoria ECONO-FJN**

Como é sabido, a economia do Nordeste historicamente distanciou-se em termos de renda per capita em relação à média do País. Concentrando quase 30% da população do Brasil, sua participação no PIB nacional não passava de cerca de 16,0% em 1998. Consequentemente, o PIB per capita do Nordeste em 1998 atingia apenas o patamar de 56,1% do respectivo PIB nacional (SUDENE, 1999). Além disso, a desvantagem no produto per capita é acrescida por uma distribuição de renda regional ainda pior que a média do país. Efetivamente, enquanto em 1997 o coeficiente de Gini era de 0,58 para o Brasil, no Nordeste chegava a 0,60 (no Sul era de 0,54). Esses indicadores, portanto, resumem uma situação econômica muito pouco satisfatória, em termos de renda per capita e de distribuição, mesmo se comparada com a também insatisfatória condição média da economia brasileira. Não obstante, vale observar que os parâmetros básicos da economia nordestina vêm, embora lentamente e aquém do desejável, apresentando melhoras. Nas últimas décadas a participação do PIB per capita do Nordeste no PIB per capita do Brasil cresceu de 41,8% para 56,1% entre 1960 e 1999. Em parte, pelo menos, isto resulta de um crescimento econômico relativamente mais rápido experimentado pela economia nordestina nas últimas quatro décadas (Tabela 1). Na verdade, o maior crescimento foi observado no setor terciário, onde localizam-se quase dois terços do PIB regional. É também verdade que esse peso maior do setor serviços resulta de um conjunto bastante heterogêneo de atividades, onde estão presentes níveis diferentes de produtividade, muitas vezes dentro do mesmo segmento. Embora a heterogeneidade seja também característica das economias menos desenvolvidas, no caso do Nordeste, periferia da periferia, ela se acentua e se difunde com os processos de crescimento e modernização que atingem de forma parcial e bastante diferenciada espaços e segmentos produtivos regionais. Nos últimos anos, com a abertura da economia brasileira essas diferenciações tendem a se exacerbar. Apesar da heterogeneidade, vale observar que o mercado nordestino é significativo e o total do seu PIB, cerca de R\$ 124,904 bilhões em 1998 (Sudene/1999) coloca a economia Nordestina à frente dos PIBs dos países da América Latina e Caribe, com exceção do México e Argentina, conforme dados

disponíveis na homepage da CEPAL para 1998. Em que pese os problemas e dificuldades, relativamente conhecidas, a economia nordestina tem demonstrado nas últimas décadas ser possuidora de algum dinamismo endógeno. O fato de ter crescido à frente da economia brasileira ao longo do período 1960-99, mostra essa capacidade de crescimento, tanto em períodos de alto quanto de baixo crescimento do conjunto da economia nacional. Ao longo de três décadas, o crescimento experimenta forte desaceleração, fato bastante conhecido e associado às mudanças mais gerais da economia brasileira. Setorialmente, a agropecuária experimenta a maior mudança, saindo de uma taxa de crescimento de 5,4% ao ano entre 1970 e 1980 para -1,6% ao ano nos anos 1990/99 (1,5% ao ano nos anos 1980/90). Cabe aqui observar que, em parte, pelo menos, esse desempenho negativo resulta de reveses climáticos observados nos anos 1990 em função do fenômeno El Niño, principalmente em 1993/94 e 1998/99. O setor industrial, que cresceu a 9,1% ao ano nos anos 1970, reduz abruptamente seu patamar de crescimento nos anos 1980/90 (1,0% ao ano), porém recupera parte do vigor nos anos 1990/99 (4,0% ao ano). No setor serviços a queda pós anos 70 é menos pronunciada e progressiva, saindo de 10,2% ao ano nos anos 70 para 3,5% ao ano nos anos 90. O exame do ocorrido em nível dos segmentos constitutivos dos grandes setores é talvez de mais utilidade para se entender onde tem se concentrado o dinamismo na região. Os dados disponibilizados pelo Departamento de Contas Regionais da SUDENE incluem alguma desagregação nos setores secundário e terciário. Na indústria pode ser observado um padrão mais regular de comportamento, tendo os segmentos energia elétrica e abastecimento de água, em grupo ou individualmente, bem como a construção liderado os índices de crescimento. A indústria extrativa mineral cresceu significativamente (6% ao ano) no período 1980/90, em boa parte em função das descobertas de petróleo e gás natural em estados como Sergipe, Alagoas e Rio Grande do Norte. A indústria de transformação no seu conjunto foi dinâmica enquanto o País exibia altas taxas de crescimento, experimentando crescimento médio negativo (-1,6% ao ano) nos anos 1980/90 e uma pequena recuperação (1,2% ao ano) no período 1990/99. No conjunto do setor industrial ao longo dos últimos trinta anos, então, os segmentos mais dinâmicos foram energia elétrica e construção, talvez por contarem em sua dinâmica com a ação, mesmo não exclusiva, de investimentos governamentais. Quanto à indústria de transformação infelizmente os dados disponíveis não estão desagregados por ramos para que se possa observar a sua dinâmica com mais propriedade. Cabe aqui apenas uma observação. No período 1981-98 a indústria de transformação no Nordeste cresceu a 1,1% ao ano em média, taxa semelhante à observada para o conjunto da economia brasileira (1,2% ao ano). Destaque-se aqui o crescimento de 2,1% ao ano observado no estado da Bahia, bastante superior, por exemplo, ao verificado em São Paulo (0,8%) no mesmo período (SUDENE, 1999). No setor terciário observa-se uma maior flutuação ou alternância de segmentos líderes do crescimento. No período 1970/80 sobressaíram-se os ramos comércio (inclusive restaurantes e hospedagem), transporte aéreo, atividades financeiras e bens imóveis. Nos anos 1980/90 os segmentos que cresceram com maior vigor foram transportes, armazenagem e comunicações, particularmente comunicações e transporte aéreo e ainda bens imóveis. Nos anos 1990/99 mantém-se esse perfil de crescimento setorial, com destaque especialmente no segmento comunicações. Ao longo das três últimas décadas, portanto, as comunicações suplantam o comércio (incluindo restaurantes e hospedagem) como atividade terciária mais dinâmica na região. A dinâmica econômica nordestina, embora com alguma variação e diversificação, devidas a

investimentos do setor privado, apresenta uma associação forte, padrão historicamente observado, vale ressaltar, com os investimentos do setor público, principalmente nos períodos de menor crescimento do PIB. Efetivamente, observando os dados relativos à formação bruta de capital fixo (FBKF) para o Brasil e para a Região nota-se que nos anos 1970 ambos experimentaram um crescimento médio de 10,2% ao ano. Nas décadas seguintes o comportamento é discrepante: nos anos 1980/90 o Nordeste amplia em 1,6% ao ano em média a FBKF enquanto no Brasil esta média é negativa (-2,6% ao ano). No período 1990/95 ocorre uma inversão, tendo o Nordeste apresentado uma taxa negativa (-1,7%) e o Brasil como um todo uma recuperação expressiva da FBKF (4,2%) (Tabela 4). Ao se observar no caso do Nordeste a desagregação da FBKF entre os setores públicos e privado nota-se o referido peso do setor público. Nos anos 1970/80, quando a economia (Nordeste e Brasil) crescia a taxas elevadas, os setores público e privado ampliaram significativamente sua FBKF. Já no período 1980/90, época de estagnação, o setor privado do Nordeste ampliou sua FBKF em média a 0,3% ao ano, enquanto o setor público apresentou uma taxa média de 3,1%. Nos anos 1990/95, em que a FBKF do nordeste decresceu a -1,7% ao ano, o setor privado cresce a 2,4% ao ano, mas o setor público apresenta uma queda de -6,3% ao ano (Tabela 5). Assim, mesmo com a recuperação dos investimentos do setor privado na primeira metade dos anos 90, o declínio acentuado dos investimentos públicos deu o tom e arrastou a FBKF da região para uma taxa negativa. A permanência da crise das finanças do setor público e de sua diminuta capacidade de crescimento é, portanto, preocupante do ponto de vista da expansão futura da economia nordestina numa conjuntura em que a economia como um todo mantém-se submetida a um regime de crescimento muito abaixo, tanto do necessário para a melhoria das condições de vida, quanto dos padrões históricos no país. Um outro aspecto que vale aqui ressaltar, para melhor caracterizar a economia nordestina, é que o Nordeste apresenta coeficientes de exportação menores que os do País como um todo. Enquanto as exportações totais representaram 8,6% do PIB brasileiro em 1999, as vendas do Nordeste para o exterior corresponderam a apenas 3,5% do PIB regional. Em que pese esta menor abertura, há que se considerar que a participação dos produtos manufaturados e semimanufaturados nas exportações nordestinas vem crescendo. Com isso, agregou-se à pauta tradicional (açúcar principalmente) uma nova lista de produtos exportáveis. Entre estes encontram-se produtos petroquímicos (pólo de Camaçari), alumínio e uma gama diversificada de produtos manufaturados (confeções, produtos alimentícios, produtos do setor metal-mecânico, etc.). Isto indica que ocorreu um certo esforço de agregação de valor aos bens produzidos para exportação, exigindo modernização e incorporação de tecnologia para manterem-se competitivos no futuro. Além disso, deve-se reiterar a necessidade de desenvolvimento de novas especializações competitivas a partir do Nordeste. O fraco desempenho exportador da região e a sua perda de importância no contexto das exportações nacionais revela um risco de exclusão a ser evitado. Por isso, a relevância da busca de uma nova articulação entre as atividades regionais de C & T e o desenvolvimento de atividades competitivas baseadas em processos locais/regionais de aprendizado e inovação.

- **Título: Getúlio Vargas e o estado nacional desenvolvimentista - Categoria CPOLI-FJN**

Quem quer que busque entender a política econômica brasileira de 1930 a 1964 não pode prescindir da figura de Getúlio Vargas, o presidente que mais tempo permaneceu à frente do governo brasileiro, seja na qualidade de líder revolucionário (1930-37), ditador (1937-45), durante o Estado Novo, e presidente constitucional (1950-54). São 19 anos ao todo, enquanto prócer máximo do País. Neste longo período, temos momentos bastante distintos e particulares na economia mundial: de 1930 a 1945, um período de crise sistêmica, a partir do crack de Wall Street, que repercutiu em todo o mundo e que no Brasil simboliza o fim da hegemonia dos cafeicultores e da política Café com Leite, bem como dos setores oligárquicos da terra, principalmente no Sudeste do País. Nos Estados Unidos da América, a política do New Deal, implantada por Roosevelt, foi uma política keynesiana antes de Keynes, cujo similar no Brasil foi a política de sustentação dos preços do café e sua queima posterior. Este processo acabou por possibilitar os primeiros passos da industrialização brasileira, que teve como impulsos fundamentais a elevação dos preços das importações e, justamente, a manutenção da demanda agregada interna, graças às compras pelo Estado do café excedente. O início da Segunda Guerra Mundial estimulou ainda mais o processo de substituição de importações, com o Estado Varguista intervindo cada vez mais a favor da industrialização. Esta política se expressa na configuração do Salário Mínimo e do Sistema Previdenciário, além da estruturação da máquina burocrática estatal. A produção industrial cresce então mais rapidamente, quase duas vezes, que a produção agrícola. O segundo período analisado (1945-1954), onde Getúlio é personagem central, apresenta características radicalmente novas no cenário internacional. Os EUA já substituem a Grã-Bretanha como nova potência mundial, apoiados na nova configuração dos organismos de Bretton Woods – o FMI, o Banco Mundial e o GATT -, instrumentos poderosos de regulação sobre o mundo; na transferência de seus grandes grupos industriais pelo mundo e na consolidação de uma força militar que garante a aplicação da pax americana, exercida até hoje a partir da persuasão armada quando o comércio não é suficientemente capaz de impô-la. O Brasil terá que aguardar a passagem medíocre do governo Dutra na Presidência, para viver o período de profundo dinamismo de Vargas, com a definição e consolidação da indústria de base, do petróleo, da energia e da química. A indústria leve de bens de consumo já fora instalada pelos nacionais, além das de matérias-primas e bens de capital. O Estado produzia aço a partir de Volta Redonda - expressão da astúcia e estratégia getulista. Nos faltava a indústria de bens de consumo durável e a ampliação das indústrias de matérias primas e de máquinas. O chamado Estado Populista passa a ser a expressão de uma ampla aliança de classes, composta pela burguesia industrial nascente, das camadas médias burocráticas e dos setores do proletariado urbano, além das oligarquias rurais não ligadas à exportação. A criação do Banco Nacional de Desenvolvimento Econômico simboliza o papel do Estado Nacional Desenvolvimentista, através dos grandes investimentos estratégicos, estruturando as bases para a industrialização do País. Não se pode tratar corretamente o nacional-desenvolvimentismo sem apontar as duas linhas ideológicas que então polarizam a intelectualidade brasileira: a corrente liberal, que se baseava na vocação agrária do País, e a corrente desenvolvimentista, que pregava a intervenção estatal para implementar a industrialização. Esta última corrente teve em Celso Furtado e Ignácio Rangel os seus maiores expoentes, que enfatizavam o caráter monopolista da estrutura industrial brasileira, absorvendo pouca mão-de-obra, o que contrastava com a disponibilidade de recursos locais, impossibilitando uma dinâmica auto-sustentada do crescimento. Os comunistas do PCB, por sua vez, indicavam a estrutura agrária

semifeudal como impeditivo maior do desenvolvimento, ao passo que Caio Prado Jr e Ruy Mauro Marini tinham na exploração imperialista o impasse fundamental. A teoria do desenvolvimentismo fora elaborada a partir da Comissão Econômica para a América Latina – CEPAL e tinha como conceitos-chaves as noções de Centro e de Periferia e de deterioração dos termos de intercâmbio ou trocas desiguais, que obstavam a industrialização da periferia. Posteriormente, estas teses foram aprofundadas no ISEB, que reuniu os principais intelectuais progressistas brasileiros e sinalizava para uma vertente mais nacionalista do desenvolvimento. O pensamento desenvolvimentista já estava presente em discurso de Vargas em 1944, quando afirmou: “Só as mentalidades impermeáveis aos ensinamentos dos fatos podem acreditar ainda na validade do ‘laissez-faire’ econômico e nos seus corolários políticos”. A ideologia desenvolvimentista populista institucionalizava o poder político das forças populares, incorporando a emergência das massas no jogo político nacional. Não houve espaço aqui para tratar de aspectos negativos de Vargas, sobretudo na sua fase Estado-Novista, onde o autoritarismo e o caráter de conciliação de classes na figura do “povo”, prevaleceram, mais ainda assim garantia o Salário Mínimo e a Consolidação das Leis do Trabalho – CLT. Muita água passou debaixo da ponte da história política brasileira desde então, bem como do capitalismo, que abandonou o crescimento virtuoso do keynesianismo, do pleno emprego e da regulação estatal, para implementar uma política neoliberal que tem gerado concentração de renda em escala mundial, privilegiando o setor rentista do capital, consolidando o desemprego estrutural e aprofundando os desníveis tecnológicos e de bem estar entre os países ricos e aqueles da periferia do sistema. Os Estados Unidos da América têm ampliado seu poder em escala mundial e pretendem firmar-se enquanto um império. O Brasil de hoje, passados 50 anos do suicídio de Vargas, se vê vulnerável na dependência dos capitais externos e submetido às exigências oriundas em grande medida dos EUA. Não tem um projeto alternativo para enfrentar os desafios das desigualdades sociais e da concentração de renda, não tem perspectiva de desenvolvimento que seja inclusivo, e teima em seguir os ditames de políticas de ajuste formuladas desde os organismos internacionais. O País tem se modernizado mas, ao mesmo tempo, excluído os trabalhadores urbanos e rurais dos frutos do bem estar. É necessário, portanto, repensar criticamente os descaminhos político-econômicos, sociais e ambientais que vêm dramatizando e esgarçando o tecido social da nação. Não basta crescimento econômico, mas um desenvolvimento que seja estruturado no bem estar dos trabalhadores e, neste sentido, o legado de Vargas tem atualidade e ensinamentos ainda hoje válidos.

B.5 Documentos RDF/XML

Para os dois textos listados acima, abaixo se encontram suas representações em RDF/XML produzidos automaticamente pelo Modelo utilizando-se o algoritmo Porter, com taxa de compressão de 70%, DMS igual a zero.

Traços gerais do desenvolvimento recente do Nordeste - Categoria ECONO-FJN

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<?xml-stylesheet href="modelo.xsl" type="text/xsl"?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
        xmlns:dc='http://purl.org/dc/elements/1.1/'
```

```

        xmlns:dcterms='http://purl.org/dc/terms/'>
<rdf:Description>
  <dc:subject>
    <rdf:Bag>
      <rdf:li>Economia</rdf:li>
      <rdf:li>ativ</rdf:li>
      <rdf:li>comunic</rdf:li>
      <rdf:li>cresciment</rdf:li>
      <rdf:li>dinâm</rdf:li>
      <rdf:li>econom brasileiro</rdf:li>
      <rdf:li>econom nordestin</rdf:li>
      <rdf:li>export</rdf:li>
      <rdf:li>fbkf</rdf:li>
      <rdf:li>indústr</rdf:li>
      <rdf:li>long</rdf:li>
      <rdf:li>méd</rdf:li>
      <rdf:li>nord</rdf:li>
      <rdf:li>país</rdf:li>
      <rdf:li>produt</rdf:li>
      <rdf:li>região</rdf:li>
      <rdf:li>setor públic</rdf:li>
    </rdf:Bag>
  </dc:subject>
  <dc:description>
    Consequentemente, o PIB per capita do Nordeste em 1998
    atingia apenas o patamar de 56, 1% do respectivo PIB nacional(
    SUDENE, 1999).
    Além disso, a desvantagem no produto per capita é acrescida
    por uma distribuição de renda regional ainda pior que a média do
    país.
    Efetivamente, enquanto em 1997 o coeficiente de Gini era de
    0, 58 para o Brasil, no Nordeste chegava a 0, 60( no Sul era de 0,
    54).
    Na verdade, o maior crescimento foi observado no setor
    terciário, onde localizam - se quase dois terços do PIB regional.
    É também verdade que esse peso maior do setor serviços
    resulta de um conjunto bastante heterogêneo de atividades, onde
    estão presentes níveis diferentes de produtividade, muitas vezes
    dentro do mesmo segmento.
    Embora a heterogeneidade seja também característica das
    economias menos desenvolvidas, no caso do Nordeste, periferia da
    periferia, ela se acentua e se difunde com os processos de
    crescimento e modernização que atingem de forma parcial e bastante
    diferenciada espaços e segmentos produtivos regionais.
    Em que pese esta menor abertura, há que se considerar que a
    participação dos produtos manufaturados e semimanufaturados nas
    exportações nordestinas vem crescendo.
    Com isso, agregou - se à pauta tradicional( açúcar
    principalmente) uma nova lista de produtos exportáveis.
    Entre estes encontram - se produtos petroquímicos( pólo de
    Camaçari), alumínio e uma gama diversificada de produtos
    manufaturados( confecções, produtos alimentícios, produtos do setor
    metal - mecânico, etc.
  </dc:description>
  <dc:publisher>FJN</dc:publisher>
  <dc:date>2006-01-17</dc:date>

```

```

<dc:type>text/xml</dc:type>
<dc:format>text</dc:format>
<dc:identifier>ECONO54</dc:identifier>
<dc:language>pt-br</dc:language>
<dc:rights>FJN</dc:rights>
<dc:title></dc:title>
<dc:author></dc:author>
<dc:contributor></dc:contributor>
<dc:relation></dc:relation>
<dc:source></dc:source>
<dc:coverage></dc:coverage>
</rdf:Description>
</rdf:RDF>

```

Título: Getúlio Vargas e o estado nacional desenvolvimentista - Categoria CPOLI-FJN

```

<?xml version='1.0' encoding='ISO-8859-1'?>
<?xml-stylesheet href="modelo.xsl" type="text/xsl"?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xmlns:dcterms='http://purl.org/dc/terms/'>
<rdf:Description>
  <dc:subject>
    <rdf:Bag>
      <rdf:li>Ciências Políticas</rdf:li>
      <rdf:li>brasil</rdf:li>
      <rdf:li>brasileir</rdf:li>
      <rdf:li>consolid</rdf:li>
      <rdf:li>desenvolviment</rdf:li>
      <rdf:li>econôm</rdf:li>
      <rdf:li>estad</rdf:li>
      <rdf:li>estrutur</rdf:li>
      <rdf:li>industrializ</rdf:li>
      <rdf:li>indústr</rdf:li>
      <rdf:li>país</rdf:li>
      <rdf:li>polít</rdf:li>
    </rdf:Bag>
  </dc:subject>
</dc:description>

```

GETÚLIO VARGAS E O ESTADO NACIONAL DESENVOLVIMENTISTA

Quem quer que busque entender a política econômica brasileira de 1930 a 1964 não pode prescindir da figura de Getúlio Vargas, o presidente que mais tempo permaneceu à frente do governo brasileiro, seja na qualidade de líder revolucionário(1930 - 37), ditador(1937 - 45), durante o Estado Novo, e presidente constitucional(1950 - 54).

Neste longo período, temos momentos bastante distintos e particulares na economia mundial: de 1930 a 1945, um período de crise sistêmica, a partir do crack de Wall Steet, que repercutiu em todo o mundo e que no Brasil simboliza o fim da hegemonia dos cafeicultores e da política Café com Leite, bem como dos setores oligárquicos da terra, principalmente no Sudeste do País.

Nos Estados Unidos da América, a política do New Deal, implantada por Roosevelt, foi uma política keynesiana antes de

Keynes, cujo similar no Brasil foi a política de sustentação dos preços do café e sua queima posterior.

A criação do Banco Nacional de Desenvolvimento Econômico simboliza o papel do Estado Nacional Desenvolvimentista, através dos grandes investimentos estratégicos, estruturando as bases para a industrialização do País.

Não se pode tratar corretamente o nacional - desenvolvimentismo sem apontar as duas linhas ideológicas que então polarizam a intelectualidade brasileira: a corrente liberal, que se baseava na vocação agrária do País, e a corrente desenvolvimentista, que pregava a intervenção estatal para implementar a industrialização.

A ideologia desenvolvimentista populista institucionalizava o poder político das forças populares, incorporando a emergência das massas no jogo político nacional.

Muita água passou debaixo da ponte da história política brasileira desde então, bem como do capitalismo, que abandonou o crescimento virtuoso do keynesianismo, do pleno emprego e da regulação estatal, para implementar uma política neoliberal que tem gerado concentração de renda em escala mundial, privilegiando o setor rentista do capital, consolidando o desemprego estrutural e aprofundando os desníveis tecnológicos e de bem estar entre os países ricos e aqueles da periferia do sistema.

```
</dc:description>
<dc:publisher>FJN</dc:publisher>
<dc:date>2006-01-17</dc:date>
<dc:type>text/xml</dc:type>
<dc:format>text</dc:format>
<dc:identifier>CPOLI133</dc:identifier>
<dc:language>pt-br</dc:language>
<dc:rights>FJN</dc:rights>
<dc:title></dc:title>
<dc:author></dc:author>
<dc:contributor></dc:contributor>
<dc:relation></dc:relation>
<dc:source></dc:source>
<dc:coverage></dc:coverage>
</rdf:Description>
</rdf:RDF>
```

B.6 Documento XSL

Abaixo encontra-se o documento XSL (modelo.xsl) definido para exibição na *Web* dos dados contidos nos documentos RDF/XML:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/"
  xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
```

```
<xsl:output method="xml" version="1.0" encoding="utf-8" omit-xml-
  declaration="yes"/>

<xsl:template match="/">
  <html>
    <head>
      <xsl:apply-templates select="rdf:RDF"
mode="metatags"/>
    </head>

    <style type="text/css">
      BODY {
        color:                black;
        font-family:          "Arial";
        font-size:             12;
      }

      .titulo {
        color:                black;
        font-family:          "Arial";
        font-size:             14;
        font-weight:          bold;
        vertical-align:        super;
      }

      .valor {
        color:                blue;
        font-family:          "Arial";
        font-size:             12;
        font-weight:          normal;
      }
    </style>

    <body>
      <xsl:apply-templates select="rdf:RDF" mode="body"/>
    </body>
  </html>
</xsl:template>

<xsl:template match="rdf:RDF" mode="metatags">
  <!-- <meta name="DC.Title"><xsl:attribute
name="content"><xsl:value-of
select="dc:title"/></xsl:attribute></meta> !-->
</xsl:template>

<xsl:template match="rdf:RDF" mode="body">
  <table width="100%" border="1" bordercolor="black">
    <tr>
      <td width="40%">
        <span class="titulo">Palavras-chave
[dc:subject]:</span>
      </td>
    </tr>

    <tr>
```

```

        <td width="60%">
            <span class="valor">
                * <xsl:for-each
select="dc:subject/rdf:Bag/rdf:li">
                    <xsl:value-of select="."/> *
                </xsl:for-each>
            </span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">Sumário
[dc:description]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:description"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:publisher]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:publisher"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:date]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:date"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:type]:</span>
        </td>
    </tr>

```

```
<tr>
  <td width="60%">
    <span class="valor"><xsl:value-of
select="dc:type"/></span>
  </td>
</tr>

<tr>
  <td width="40%">
    <span class="titulo">[dc:format]:</span>
  </td>
</tr>

<tr>
  <td width="60%">
    <span class="valor"><xsl:value-of
select="dc:format"/></span>
  </td>
</tr>

<tr>
  <td width="40%">
    <span class="titulo">[dc:identifier]:</span>
  </td>
</tr>

<tr>
  <td width="60%">
    <span class="valor"><xsl:value-of
select="dc:identifier"/></span>
  </td>
</tr>

<tr>
  <td width="40%">
    <span class="titulo">[dc:language]:</span>
  </td>
</tr>

<tr>
  <td width="60%">
    <span class="valor"><xsl:value-of
select="dc:language"/></span>
  </td>
</tr>

<tr>
  <td width="40%">
    <span class="titulo">[dc:rights]:</span>
  </td>
</tr>

<tr>
  <td width="60%">
    <span class="valor"><xsl:value-of
select="dc:rights"/></span>
```

```

        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:title]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:title"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:author]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:author"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:contributor]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:contributor"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">
            <span class="titulo">[dc:relation]:</span>
        </td>
    </tr>

    <tr>
        <td width="60%">
            <span class="valor"><xsl:value-of
select="dc:relation"/></span>
        </td>
    </tr>

    <tr>
        <td width="40%">

```

```

                <span class="titulo">[dc:source]:</span>
            </td>
        </tr>

        <tr>
            <td width="60%">
                <span class="valor"><xsl:value-of
select="dc:source"/></span>
            </td>
        </tr>

        <tr>
            <td width="40%">
                <span class="titulo">[dc:coverage]:</span>
            </td>
        </tr>

        <tr>
            <td width="60%">
                <span class="valor"><xsl:value-of
select="dc:coverage"/></span>
            </td>
        </tr>
    </table>
</xsl:template>
</xsl:stylesheet>

```

B.7 Página Web

Abaixo encontra-se a exibição na *Web* do documento RDF/XML “Getúlio Vargas e o estado nacional desenvolvimentista”, quando aplicado a XSL definida acima:

Palavras-chave [dc:subject]:
* Ciências Políticas * brasil * brasileiro * consolid * desenvolviment * econôm * estad * estrutur * industrializ * indústr * país * polít *
Sumário [dc:description]:
GETÚLIO VARGAS E O ESTADO NACIONAL DESENVOLVIMENTISTA Quem quer que busque entender a política econômica brasileira de 1930 a 1964 não pode prescindir da figura de Getúlio Vargas, o presidente que mais tempo permaneceu à frente do governo brasileiro, seja na qualidade de líder revolucionário(1930 - 37), ditador(1937 - 45), durante o Estado Novo, e presidente constitucional(1950 - 54). Nos Estados Unidos da América, a política do New Deal, implantada por Roosevelt, foi uma política keynesiana antes de Keynes, cujo similar no Brasil foi a política de sustentação dos preços do café e sua queima posterior. A ideologia desenvolvimentista populista institucionalizava o poder político das forças populares,

incorporando a emergência das massas no jogo político nacional.
[dc:publisher]:
FJN
[dc:date]:
2006-01-30
[dc:type]:
text
[dc:format]:
text/xml
[dc:identifier]:
CPOLI46
[dc:language]:
pt-br
[dc:rights]:
FJN
[dc:title]:
[dc:author]:
[dc:contributor]:
[dc:relation]:
[dc:source]:
[dc:coverage]:

Figura B.1 – Exibição das Informações de um Documento RDF/XML, na *Web*

Bibliografia

- AB1984 ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. Beverly Hills, CA: Sage, 1984. 88 p.
- ADL-SCORM ADL (SCORM). Advanced Distributed Learning (Sharable Content Object Reference Model). Disponível em: <<http://www.adlnet.org/scorm/index.cfm>>. Acesso em: junho 2005.
- AMEGA2005 GREENBERG, J.; SPURGIN, K.; CRYSTAL, A. **Final Report for the AMeGA (Automatic Metadata Generation Applications)**. Submitted to the Library of Congress. February 17, 2005. Disponível em: <http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf>. Acesso em: junho 2005
- Barbosa2003 BARBOSA, I. C. **Prospecção De Dados na Manutenção Semi-Automática de Ontologias**. Dissertação de mestrado: Universidade Federal de Minas Gerais – UFMG. 2003.
- Barzilay1997 BARZILAY, R. **Lexical Chains for Summarization**. Dissertação de mestrado: Ben-Gurion University of the Negev, 1997. Disponível em <<http://www.cs.bgu.ac.il/~elhadad/barzilay.pdf>>. Acesso em: set 2005.
- Baxendale1958 BAXENDALE, P.B. **Machine-made index for technical literature – an experiment**. IBM Journal of Research and Development, Vol. 2, pp. 354-365. 1958.
- BB1975 BORKO, H.; BERNIER, C. **Abstracting Concepts and Methods**. Academic Press, New York. 1975.
- Berners1998 BERNERS-LEE, T. **Metadata Architecture: Documents, Metadata, and Links. Axiom of Web Architecture**. 1997. Atualizado em 1998. Disponível em: <<http://www.w3.org/DesignIssues/Metadata.html>>. Acesso em: junho 2005.
- BJ1988 BLACK, W. J.; JOHNSON, F.C. **A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques**. Expert Systems for Information Management, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology. 1988.
- BK1997 BOGURAEV, B.; KENNEDY, C. **Salience-Based Content Characterisation of Text Documents**. In I. Mani and M. Maybury (eds.). Proc. of the Intelligent Scalable Text Summarization Workshop, pp. 2-9. ACL/EACL'97 Joint Conference. Madrid, Spain. 1997

- BLC2000 BRAGA, A.P.; LUDERMIR, T.B.; CARVALHO, A.C.P.L.F. **Redes Neurais Artificiais: Teoria e Aplicações**. LTC - Livros Técnicos e Científicos Editora S.A, Rio de Janeiro. 2000.
- BLH2001 BERNERS-LEE, T.; LASSILA, O.; HENDLER, J. **The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities**. Scientific American, Maio 2001; Disponível em: <<http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84^9809EC588EF21>>. Acesso em: outubro/2004.
- Borg1999 BORGMAN, C. L. **What are digital libraries?** Competing visions, In Information Processing and Management. Vol. 35, Nrº. 3, pp. 227-243. Janeiro de 1999.
- BR1999 BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York : ACM, 1999.
- Bullock1999 BULLOCK, A. **Preservation of Digital Information: Issues and Current Status**. National Library of Canada, 1999. Disponível em: <<http://www.collectionscanada.ca/9/1/p1-259-e.html>>. Acesso em: maio 2002.
- Cardoso2000 CARDOSO, O. N. P. **Recuperação de Informação**. Infocomp Revista de Computação da Ufla, Lavras - MG, v. 1, p. 33-38, 2000.
- CJB1999 CHANDRASEKARAN, B.; JOSEPHSON, J.R.; BENJAMINS, V.R. **What are Ontologies, and Why do We Need Them?**. IEEE Intelligent Systems, p. 20-26. Janeiro/Fevereiro de 1999.
- Cleve1998 CLEVELAND, G. **Digital Libraries: Definitions, Issues and Challenges**. UDT Core Programme, 1998. Disponível em: <www.ifla.org/VI/5/op/udtop8/udtop8.htm>. Acesso em: maio 2002.
- CMNS1996 CHEN, H.; MARTINEZ, J.; NG, T. D.; SCHATZ, B. R. **A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An experiment on the Worm Community System**. Journal of the American Society for Information Science, Volume 47, Number 8, August 1996. Disponível em: <<http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html#272>>. Acesso em: março 2005.
- Columbia2001 COLUMBIA UNIVERSITY LIBRARIES. Selection Criteria For Digital Imaging. Disponível em: <<http://www.columbia.edu/cu/libraries/digital/criteria.html>>. Acesso em: abril 2002.
- Congress1 LIBRARY OF CONGRESS. Challenges to Building an Effective Digital Library . Disponível em: <<http://memory.loc.gov/ammem/dli2/html/cbedl.html>>. Acesso em: maio 2002.

- Congress2 LIBRARY OF CONGRESS. The Mission and Strategic Priorities of the Library of Congress. 1997-2004. Disponível em: <<http://lcweb.loc.gov/ndl/mission.html>>. Acesso em: maio 2002.
- Cremmins1996 CREMMINS, E. T. **The Art of Abstracting**. Information Resources Press, Arlington, VA, second edition. 1996.
- DAML DAML+OIL. DARPA Agent Markup Language – Ontology Interchange Language. Disponível em: <<http://www.w3.org/TR/daml+oil-reference>>. Acesso em: junho 2005.
- DublinCore DUBLIN CORE METADATA INITIATIVE. Disponível em: <<http://dublincore.org>>. Acesso em: agosto 2004.
- Edmund1969 EDMUNDSON, H.P. **New Methods in Automatic Extracting**. *Journal of the ACM*, 16, pp. 264-285. 1969.
- ER2002 ESPINA, A. P; RINO, L. H. M. **Utilização de Métodos Extrativos na Sumarização**. Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional - NILC-TR-02-06. Março, 2002. Disponível em: <<http://www.dc.ufscar.br/~lucia/TechRep/NILCTR0206-EspinaRino.pdf>>. Acesso em: maio 2004.
- Ethnologue ETHNOLOGUE. Disponível em: <<http://www.ethnologue.com>>. Acesso em: julho 2005.
- FGDC FGDC. Content Standard for Digital Geospatial Metadata. Disponível em: <<http://www.fgdc.gov/metadata/constan.html>>. Acesso em: junho 2005.
- FTFD1997 FEDERAL TASK FORCE ON DIGITIZATION (CANADÁ). **Towards a Learning Nation: The Digital Contribution - Final Report**. 1997. Disponível em: <<http://www.nlc-bnc.ca/8/3/r3-407-e.html>>. Acesso em: agosto 2002.
- Garcia2005 GARCIA, E. **The Extended Boolean Model Weighted Queries**. Article 6 of the series *Term Vector Theory and Keyword Weights*. 2005. Disponível em: <<http://www.miislita.com/term-vector/term-vector-6-boolean-model.html>>. Acesso em: julho 2005.
- GF2001 GONÇALVES, M.A.; FOX, E.A.; **Technology and Research in a Global Networked**. University Digital Library (NUDL). Digital Library Research Laboratory at Virginia. 2001.
- Gilliland2000 GILLILAND-SWETLAND, A. J. **Introduction to Metadata: Setting the Stage**, 2000. Disponível em: <http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html>. Acesso em: ag. 2000.

- GILS GILS. Global Information Locator Service. Disponível em: <<http://www.gils.net/>>. Acesso em: junho 2005.
- Greenberg2003 GREENBERG, J. **Metadata and the World Wide Web**. In M.S. Drake (Ed.) *Encyclopedia of library and information science* (2nd ed.) (pp.1876-1888). New York: Marcel Dekker, Inc. 2003.
- Harvard2004 HARVARD UNIVERSITY LIBRARY. **Selection for digitization. A decision-making matrix**. 2004. Disponível em: <<http://preserve.harvard.edu/bibliographies/matrix.pdf>>. Acesso em: agosto 2005.
- HL1999 HOVY, E.; LIN, C. **Automated text summarization in SUMMARIST**. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 81–94. 1999.
- HM2002 HOWARD, D. W.; MICHAEL, G. C. **Digital Video Archives: Managing through Metadata**. In Building a National Strategy for Digital Preservation. National Digital Information Infrastructure and Preservation Program, Library of Congress. April, 2002. Disponível em: <<http://www.informedia.cs.cmu.edu/documents/Wactlar-CLIR-final.pdf>>. Acesso em: maio 2004.
- Hodge2000 HODGE, G.M. **Best Practices for Digital Archiving - An Information Life Cycle Approach**. D-Lib Magazine, January 2000, Volume 6, Number 1. 2000.
- Hodgson1998 HODGSON, K. **Metadata: Foundations, Potential and Applications**. School of Library and Information Studies. University of Alberta, Canada. Disponível em: <<http://www.slis.ualberta.ca/538/khodgson/metadata.htm>>. Acesso em: Abril 2002.
- HR1998 HEMINGER, A. R.; ROBERTSON, S. B. **Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents**. November 21, 1998. Disponível em: <<http://www.ercim.org/publication/ws-proceedings/DELOS6/rosetta.pdf>>. Acesso em: maio 2004.
- ISC INTERNET SYSTEMS CONSORTIUM. Disponível em: <<http://www.isc.org/>>. Acesso em: fev. 2003.
- ITESM1999 ITESM. Biblioteca Digital del Tecnológico de Monterrey. 1999. Disponível em: <http://cib.mty.itesm.mx/bd/working_paper.pdf>. Acessado em: out 2000.
- Jacob2003 JACOB, E. K. **Ontologies and the Semantic Web**. In Special Section of Bulletin of the American Society for Information Science and Technology. Maio 2003. Disponível em: <<http://www.asis.org/Bulletin/Apr-03/jacob.html>>. Acesso em: jun 2005.
- JJBW1998 JENKINS. C.; JACKSON, M.; BURDEN, P.; WALLIS, J. **Automatic RDF**

- Metadata Generation for Resource Discovery.** Disponível em: <http://www.scit.wlv.ac.uk/~ex1253/rdf_paper/>. Acesso em: agosto 2003.
- JJBW1998-a JENKINS, C.; JACKSON, M.; BURDEN, P.; WALLIS, J. **Automatic Classification of Web Resources using Java and Dewey Decimal Classification.** Computer Networks and ISDN Systems, Volume 30 646-648, 1998.
- KPC1995 KUPIEC, J.; PETERSEN, J.; CHEN, F. **A trainable document summarizer.** In Edward Fox, Peter Ingwersen, & Raya Fidel (eds.). *Proceedings of the 18th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 68-73, Seattle, WA, EUA. 1995.
- Lei1998 LEI N.º 9.610, DE 19 DE FEVEREIRO DE 1998: ALTERA, ATUALIZA E CONSOLIDA A LEGISLAÇÃO SOBRE DIREITOS AUTORAIS E DÁ OUTRAS PROVIDÊNCIAS. Disponível em: <http://www.mct.gov.br/legis/leis/9610_98.htm>. Acesso em: maio 2002.
- LJ1996 LEWIS, D.D.; JONES, K. S. **Natural language processing for information retrieval.** Communications of the ACM; Volume 39 , 92 – 101p. Janeiro 1996.
- Lopes2004 LOPES, M. C.S. **Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português.** Tese - Universidade Federal do Rio de Janeiro; 2004.
- LR2001 LOH, S.; GARIN, R. S. **Web Intelligence: Inteligência Artificial para Descoberta de Conhecimento na Web.** V Oficina de Inteligência Artificial, pp.11-34. Universidade Católica de Pelotas, Editora EDUCAT, Pelotas-RS, 22 de Novembro de 2001.
- LSKF2000 LARocca NETO, J.; SANTOS, A.D.; KAESTNER, A.A.; FREITAS, A.A. **Generating Text Summaries through the Relative Importance of Topics.** In the Proceedings of the International Joint Conference IBERAMIA/SBIA, Atibaia, SP. 2000.
- LTSC-IMS LTSC. Learning Technology Standards Committee (Instructional Management System). Disponível em: <<http://www.learnativity.com/standards.html>>. Acesso em: maio 2005.
- Luhn1958 LUHN, H. P. **The automatic creation of literature abstracts.** IBM Journal of Research and Development, Vol. 2, pp. 159-165. 1958.
- LWO2000 LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. **Concept-based Knowledge Discovery in Texts Extracted from the Web.** ACM SIGKDD Explorations, v.2, n.1, July 2000, pp.29-39.
- MANI2001 MANI, I. **Summarization Evaluation: An Overview.** In the *Proceedings of the Workshop on Automatic Summarization*. Pittsburgh, Pennsylvania. 2001. Também

- disponível em: <<http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/summani.pdf>>. Acesso em maio 2005.
- MARC AACR2/MARC. Disponível em: <<http://www.loc.gov/marc/>>. Acesso em junho 2005.
- Marco2000 MARCO, D. **Building and Managing the Meta Data Repository – A Full Lifecycle Guide**. Robert M. Elliott. 2000.
- Marcu1997 MARCU, D. **From Discourse Structures to Text Summaries**. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop*, pp. 82-88. ACL/EACL'97 Joint Conference. Madrid, Spain. 1997.
- Marcu2000 MARCU, D. **The Theory and Practice of Discourse Parsing and Summarization**. The MIT Press. Cambridge, Massachusetts. 2000.
- MH1991 MORRIS, J.; Hirst, G. **Lexical cohesion, the thesaurus, and the structure of text**. *Computational Linguistics*, 17(1): 21-48.
- MM2004 MANOLA, F.; MILLER, E. **RDF Primer**. W3C Recommendation February 2004. Disponível em: <<http://www.w3.org/TR/rdf-primer/>>. Acesso em: abril 2005.
- MPER2001 MARTINS, C.B.; PARDO, T.A.S.; ESPINA, A.; RINO, L.H.M. **Introdução à Sumarização Automática**. Relatório Técnico do Departamento de Computação da UFSCar. Disponível em: <<http://www.nilc.icmc.usp.br/~thiago/RTDC00201-CMartinsEtAl.pdf>>. Acessado em: abril 2005.
- MT1988 MANN, W.C.; THOMPSON, S.A. **Rhetorical Structure Theory: Toward a Functional Theory of Text Organization**. *Text*, 8 (3), pp. 243-281.
- Nwana1996 NWANA, H. **Software agents: an overview**. *Knowledge Engineering Review*, v. 11, n. 3. p. 205-244, 1996.
- OH2001 ORENGO, V. M.; HUYCK, C.R. **A Stemming Algorithm for The Portuguese Language**. In *Proceedings of the SPIRE Conference*, Laguna de San Raphael, Chile. November 13-15, 2001.
- OIL OIL. *Ontology Interchange Language*; Disponível em: <<http://xml.coverpages.org/oil.html>>. Acesso em: maio 2005.
- Orengo2004 ORENGO, V. M. **Assessing Relevance Using Automatically Translated Documents for Cross-Language Information Retrieval**. Tese apresentada a Middlesex University. 2004.
- Paice1981 PAICE, C. D. **The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases**. *Information Retrieval Research*. Butterworth & Co. (Publishers). 1981.

- Porter1980 PORTER, M. **An algorithm for suffixing stripping**. Program, Volume 14, Number 3, pages 130-137, 1980.
- PTC1999 PRETO, G.; TANAKA, K.; CAMPOS, M. L. M. **MetaSIG: Ambiente de Metadados para Aplicações de Sistemas de Informações Geográficas**. Instituto Militar de Engenharia – Depto de Engenharia de Sistemas. Setembro 1999.
- PZ1975 POLLOCK, J.J.; ZAMORA, A. **Automatic Abstracting Research at Chemical Abstracts Service**. *Journal of Chemical Information and Compute Sciences* 15(4): 226-232. 1975.
- RJB2000 RADEV, D.R.; JING, H.; BUDZIKOWSKA, M. **Summarization of multiple documents: clustering, sentence extraction, and evaluation**. In the *Proceedings of the Workshop on Automatic Summarization*, pp. 21-30. Seattle, WA. 2000.
- Rothenb1999 ROTHENBERG, J. **Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation**. Report to CLIR, January, 1999. Disponível em: <<http://www.clir.org/pubs/reports/rothenberg/contents.html>>. Acesso em: março 2004.
- RP2003 RINO, L. H. M; PARDO, T.A.S. **A Sumarização Automática de Textos: Principais Características e Metodologias**. In Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, pp. 203-245. Campinas-SP, Brasil; Agosto, 2003. Disponível em: <<http://www.nilc.icmc.usp.br/~thiago/JAIA2003-RinoPardo.pdf>>. Acesso em: março 2005.
- SA2004 SOUZA, R.R.; ALVARENGA, L. **A Web Semântica e suas contribuições para a ciência da informação**. Disponível em <http://www.scielo.br/scielo.php?pid=S0100-19652004000100016&script=sci_arttext&tlng=pt>. Acesso em: outubro 2004.
- SB1988 SALTON, G; BUCCLEY, C. **Term-weighting approaches in Automatic Retrieval**. *Information Processing & Management*, Vol. 24, No 5, 1988.
- SemanticWeb SEMANTIC WEB. Disponível em: <<http://www.semanticweb.org/about.html>>. Acesso em: out 2004.
- SEMD2000 STAAB, S.; ERDMANN, M.; MAEDCHE, A.; DECKER, S. **An Extensible Approach for Modeling Ontologies in RDF(S)**. ECDL 2000 Workshop on the Semantic Web. Disponível em: <http://www.ics.forth.gr/isl/SemWeb/proceedings/session2-1/html_version/>. Acesso em: abril 2005.
- SFW1983 SALTON, G.; FOX, E.A.; WU, H. **Extended Boolean information retrieval**. *Communications of the ACM*, 26(11):1022-1036. 1983.

- SG1996 SPARCK Jones, K.; GALLIERS, J.R. **Evaluating Natural Language Processing Systems: An Analysis and Review**. Lecture Notes in Artificial Intelligence, Vol. 1083.1996.
- SHOE SHOE. Simple HTML Ontology Extensions. Disponível em: <<http://xml.coverpages.org/shoe.html>>. Acesso em: maio 2005.
- SL2000 SAGGION, H.; LAPALME, G. **Concept identification and presentation in the context of technical text summarization**. In the Proceedings of the NAACL-ANLP Workshop on Automatic Summarization, pp. 1-10. Seattle, WA. 2000.
- SM1983 SALTON, G.; McGill, M.J. **Introduction to Modern Information Retrieval**. McGraw-Hill. New York. 1983.
- SMZ2003 SHI, R.; MALY, K.; ZUBAIR, M. **Automatic Metadata Discovery from non-cooperative Digital Libraries**. Dept. of Computer Science, Old Dominion University, USA, 2003.
- Sparck1993 SPARCK, K. **What might be in a summary?** In Krause Knorz and Womser-Hacker (eds.), *Information Retrieval 93*, pp. 9-26. Universitätsverlag Konstanz. 1993.
- Sparck1999 SPARCK, K. **Automatic summarizing: Factors and directions**. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 1–13. 1999.
- SSMB1997 SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. **Automatic Text Structuring and Summarization**. *Information Processing & Management*, 33(2), pp. 193-207. 1997.
- SVM2000 SOUZA, M.; VENDRUSCULO, L.; MELO, G. **Metadata for description of electronic information resources: the use of Dublin Core**. Ci. Inf., Brasília, v. 29, n. 1, p. 93-102, jan./abr. 2000.
- SZR2000 SRIHARI, R. K.; ZHANG, Z.; RAO, A. **Intelligent indexing and semantic retrieval of multimodal documents**. *Information Retrieval*, 2(2/3):245–275, 2000.
- Tannenb2001 TANNENBAUM, A. **Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand**. Addison Wesley, 2001.
- TM1999 TEUFEL, S.; MOENS, M. **Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting**. In Inderjeet Mani and Mark T. Maybury (Eds.), *Advances in Automatic Text Summarization*. Massachusetts Institute of Technology Press. 1999.

-
- W3C-RDF W3C. Resource Description Framework (RDF). Disponível em: <<http://www.w3.org/RDF/>>. Acesso em: março 2005.
- W3C-SW W3C. Semantic Web Activity Statement. Disponível em: <<http://www.w3.org/2001/sw/Activity.html>>. Acesso em: março 2005.
- W3C-XML W3C. Extensible Markup Language (XML). W3C Recommendation 04 February 2004. Disponível em: <<http://www.w3.org/TR/REC-xml/>>. Acesso em: março 2005.
- XOL XOL. Ontology Exchange Language. Disponível em: <<http://xml.coverpages.org/xol.html>>. Acesso em: maio 2005.
- YSLH2003 YEE, K.; SWEARINGEN, K.; LI, K.; HEARST, M. **Faceted Metadata for Image Search and Browsing**. CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA. Copyright 2003 ACM 1-58113-630-7/03/0004. Disponível em: <<http://citeseer.ist.psu.edu/cache/papers/cs/28078/http:zSzzSzsbailando.sims.berkeley.edu/zSzpaperszSzflamenco-chi03.pdf/yee03faceted.pdf>>. Acesso em: agosto 2003.