Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Uma Abordagem para a Indexação Semântica de Documentos Textuais Baseada Em Fontes Heterogêneas de Informação

José Gildo de Araújo Júnior

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Recuperação da Informação

Ulrich Schiel, Leandro Balby
(Orientadores)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

A663a Araújo Júnior, José Gildo de.

Uma abordagem para a indexação semântica de documentos textuais baseada em fontes heterogêneas de informação / José Gildo de Araújo Júnior. – Campina Grande, 2013.

99 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2013.

"Orientação: Prof. Dr. Ulrich Schiel, Prof. Dr. Leandro Balby".

Referências.

1. Sistemas de Recuperação da Informação. 2. Indexação Semântica. 3. Fontes Heterogêneas. I. Schiel, Ulrich. II. Balby, Leandro. III. Título.

CDU 004.414.28(043)

"UMA ABORDAGEM PARA INDEXAÇÃO SEMÂNTICA DE DOCUMENTOS TEXTUAIS BASEADA EM FONTES HETEROGÊNEAS DE INFORMAÇÃO"

JOSÉ GILDO DE ARAÚJO JÚNIOR

DISSERTAÇÃO APROVADA EM 19/04/2013

ULRICH SCHIEL, Dr., UFCG
Orientador(a)

LEANDRO BALBY MARINHO, Dr., UFCG
Orientador(a)

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Examinador(a)

MARIA FERNANDA MOURA, Drª, EMBRAPA
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Atualmente, um dos principais desafios no campo da Recuperação de Informação (RI) é o desenvolvimento de sistemas que processem corretamente a ideia ou conceito por trás das consultas emitidas pelos usuários. Sistemas convencionais de RI, geralmente limitam suas funcionalidades à indexação e recuperação por palavras-chave, mecanismo que gera resultados incipientes quando termos indexados não são mencionados na consulta. Consultas tais como: *"O rei da música brasileira"* e *"Roberto Carlos"*, mesmo utilizando um distinto grupo de palavras, podem representar a mesma ideia ou conceito e, portanto, o sistema deveria retornar o mesmo conjunto resposta. Entretanto, para sistemas de RI que não consideram o aspecto semântico, ambas consultas retornarão, eventualmente, conjuntos respostas distintos.

Propõe-se, neste trabalho, um novo paradigma de indexação semântica de conceitos, onde, neste novo enfoque, conceitos presentes em documentos textuais são enriquecidos semanticamente de maneira automática por meio de informações presentes em fontes heterogêneas de informação, unindo, em um único ambiente, características de dicionários, enciclopédias e de sentido comum. Desta maneira, isola-se a ideia ou conceitualização dos objetos de suas inúmeras formas de representação.

A abordagem proposta foi comparada com o projeto UBY, um recurso léxico-semântico de grande escala que combina uma vasta gama de informações construídas tanto por peritos quanto coletivamente para o idioma Inglês e Alemão. De maneira que ambas foram submetidas a diversas coleções de documentos e foi comprovada a superioridade da abordagem proposta quando comparada ao UBY. Para isso, mediu-se o número de conceitos presentes nas coleções de documentos identificados por ambas as abordagens; a conectividade, onde computou-se para cada elemento identificado o número de conexões estabelecidas com outros conceitos; e, a qualidade do enriquecimento semântico produzido, onde foram computadas as relações semânticas estabelecidas entre conceitos.

Palavras-chave: *indexação semântica, fontes heterogêneas, recuperação da informação.*

Abstract

Nowadays, one of the main challenges in the area of Information Retrieval (IR) is the development of systems that correctly process the idea or concept in the queries emitted by users. Conventional IR systems usually limit their functionality to indexing and retrieving keywords, which creates incipient results when indexed terms are not mentioned in the query. Queries such as: *"The king of Brazilian music"* and *"Roberto Carlos"*, even using a distinguished group of words, may represent the same idea or concept; therefore, the system should return the same set of answers. However, for IR systems that do not consider the semantic aspect, both queries return different answering sets.

In this work, we proposed a new paradigm of semantic indexing of concepts. With this new approach, concepts present in textual documents are semantic enriched automatically using information which is presented in heterogeneous sources joined in a single environment features of dictionaries, encyclopedias and common sense. In this way, the idea of object contextualization is isolated from the several forms of object representations.

The proposed approach was compared with UBY project, a large scale lexic-semantic resource which combines a wide range of information built by experts and collectively for English and German languages. Both approaches were subjected to various collections of documents and was proven the superiority of the proposed approach compared to UBY. To make this conclusion we measured: the number of concepts found in the collections of documents identified by either approach; connectivity, which was computed for each element identified the number of connections established with other concepts; and quality of produced semantic enrichment, which was computed if the semantic relations between concepts established are consistent.

Keywords: *semantic indexing, heterogeneous sources, information retrieval, semantic enrichment.*

Agradecimentos

Deus, Javé, Jeová, ou como definam, obrigado por me proporcionar dias de sol e chuva, saúde constante, e uma determinação inabalável que me permitiram acordar cedo, dormir tarde, e repetir o ciclo até a conclusão deste trabalho.

Ao meu pai, mamãe e familiares, tão pouco é um muito obrigado diante de tanto suporte. Obrigado por me mostrarem o caminho mais coerente em cada momento, acalmarem meu coração quando eu queria queimar etapas na vida e me receberem de braços abertos sempre que assumi a postura de filho pródigo. Acima de tudo, obrigado por acreditarem em mim, mesmo quando em alguns momentos eu mesmo já havia desistido.

Aos Doutores Ulrich Schiel e Leandro Balby, muito obrigado por tudo. Sempre foi surpreendente ver minhas teorias sendo desmoronadas por perguntas tão simples e ser motivado a reconstruí-las mais fortes e melhor. Muito obrigado pela humildade, por sempre estarem solícitos a esclarecerem uma dúvida, ou doarem generosamente 5 minutos de seu tempo, que facilmente se convertiam em horas. Para mim, sempre foi motivo de muito orgulho tentar contribuir com minha energia e disposição.

Aos Doutores Eustáquio Rangel e Cláudio Baptista pela boa vontade e prontidão sempre que necessitei. Doutores, muito obrigado.

Aos meus amigos do peito, e de todas as horas, certas ou incertas: Anderson Lucena, Wesklay(gluglu), Alfeu Buriti e Dayse Guimarães, Eric Alexandre, Adriano Santos, Isabel Nunes, Magna Celi, Vladmir Catão e tantos outros, pelo tempo dedicado, pelas discussões, gargalhadas, brincadeiras, preocupações e pelos convites de sair e espairecer.

À Patricia Alanis Maldonado, a coincidência mais formosa da minha vida, que por um momento de minha história foi meu divino complemento e que, pelas situações da vida, encontra-se longe. Jamais esquecerei do amor simples e puro que brotou quando vivíamos nos fundos de uma sorveteria, obrigado por trazer luz a muitos dos meus dias.

À minha filha, Ana Luiza Pontes de Lucena Araújo, principal fonte de motivação, energia, inspiração e dedicação, sentimento incomensurável, singelo, sublime e inexplicável, no qual todas as coisas se justificam. A você minha filha, dedico todo esforço. Obrigado por sempre compreender quando papai tinha que trabalhar. . . eu te amo.

Sumário

1 Introdução **1**

1.1 Objetivos 9

1.1.1 Objetivo Geral 9

1.1.2 Objetivos Específicos 10

1.2 Relevância 10

1.3 Escopo de Trabalho 11

2 Fundamentação Teórica **12**

2.1 Termo e Conceito 12

2.2 Representações do Conhecimento 13

2.2.1 Tesauros 13

2.2.2 Dicionários 14

2.2.3 Enciclopédia 16

2.2.4 Rede Semântica 16

2.2.5 Mapa de Tópicos 17

2.3 Relações Semânticas 19

2.4 Fontes Externas e Heterogêneas de Informação 22

2.4.1 WordNet 23

2.4.2 Reverb 24

2.4.3 Wikipédia e Wiktionary 24

2.4.4 Verbosity 25

2.5 Ontologias 26

2.6 Bibliotecas Digitais 27

2.7 ESA 27

2.7.1 Rápida Formalização do Algoritmo ESA 28
2.8 Projeto RISO-T 29
2.8.1 RISO-VTD: Criação de Vetores Temáticos de Domínios 32
2.8.2 RISO-ET: Identificação e Tradução Semântica de Conceitos Espaço-Temporais 33
2.8.3 RISO-ES: Enriquecimento Semântico de Conceitos, Desambiguação e Indexação Semântica de Documentos 33
2.8.4 RISO-PC: Processamento de Consultas e Ambiente de Interatividade 34

3 Trabalhos Relacionados 35
3.1 Retroalimentação de Relevância 35
3.2 Expansão de Consultas 36
3.3 Sistemas de RI Auxiliados por Formulários 37
3.4 Relações Semânticas Utilizando Wikipédia 38
3.5 Validação de Algoritmos de RI 38
3.6 Relacionamento de Termos da Consulta com Conceitos da Wikipédia . . . 39
3.7 Junção de Fontes de Informação 40
3.8 Indexação Semântica 41
3.9 Análise Formal de Conceitos 42
3.10 Projeto UBY 42
3.11 Desafios da RI 43
3.12 Comentários Finais 43

4 RISO-ES: Enriquecimento Semântico, Desambiguação e Indexação 44
4.1 Metodologia de Trabalho 46
4.1.1 *Primeira fase:* Obtenção de Conceitos em Documentos Textuais . . 47
4.1.2 *Segunda fase:* União de Fontes Heterogêneas de Informação 47
4.1.3 *Terceira fase:* Enriquecimento Semântico de Conceitos 48
4.1.4 *Quarta Fase:* Construção Automática de Ontologias 49
4.1.5 *Quinta Fase:* Desambiguação e Indexação de Conceitos 50
4.1.6 *Sexta fase:* Avaliação e Comparação da Ferramenta 50
4.2 Formalização do Problema 51

4.3 Enriquecimento Semântico de Conceitos 55
4.3.1 Extração de Conceitos 55
4.3.2 Enriquecimento Semântico 56
4.3.3 Construção de Ontologia 59
4.3.4 Desambiguação, Indexação Semântica e Construção de um Mapa de Tópicos Semântico 62
4.3.5 Comentários Finais 68

5 Validação e Verificação **69**
5.1 Cobertura e Conectividade 70
5.1.1 Coleção de Teste Reuters-21578 71
5.1.2 Analisadores Morfológicos 73
5.2 Qualidade do Enriquecimento Semântico Proporcionado 74
5.2.1 Enriquecimento Semântico 76
5.3 Tempo de Processamento 79
5.3.1 Tempo de Processamento para o Cálculo da Cobertura e Conectividade 81
5.3.2 Tempo de Processamento para o Cálculo do Enriquecimento Semântico 81
5.4 Considerações Finais 82

6 Conclusão e Trabalhos Futuros **85**

A Detalhes da Validação e Verificação **94**
A.1 Enriquecimento Semântico 94
A.1.1 Software ESA 94
A.1.2 Conceitos Selecionados 94
A.2 Dados Reuters-21758 98
A.2.1 Siglas e Acrônimos 98
A.2.2 Regiões Geográficas 98
A.2.3 Pessoas 98
A.3 Analisadores Morfológicos 98
A.3.1 POS 98
A.3.2 Hunspell 99

A.3.3 Agid-4 99

Lista de Símbolos

UBY - *Recurso léxico semântico para o processamento de linguagem natural*

DVS - *Decomposição de Valor Singular*

ESA - *Explicit Semantic Analysis*

LD - *Linguagem Documental*

NCI - *National Cancer Institute*

NIAAA - *National Institute on Alcohol Abuse and Alcoholism*

PLN - *Processamento de Linguagem Natural*

RI - *Recuperação de Informação*

RISO-EET - *Recuperação de Informação Semântica de Objetos - Enriquecimento Espacial Temporal*

RISO-ES - *Recuperação de Informação Semântica de Objetos - Enriquecimento Semântico*

RISO-PC - *Recuperação de Informação Semântica de Objetos - Processamento de Consulta*

RISO-T - *Recuperação de Informação Semântica de Objetos Textuais*

RISO-VTD - *Recuperação de Informação Semântica de Objetos - Vetores Temáticos de Domínio*

SGBDOR - *Sistema Gerenciador de Banco de Dados Objeto Relacional*

TREC - *Text Retrieval Conference*

Lista de Figuras

1.1 Processo de recuperação de informação não semantizado. 3
1.2 Processo de recuperação de informação semantizado. 4
1.3 Consulta conceitual com interpretação sintática. 5
1.4 Consulta conceitual com interpretação semântica. 5
1.5 Exemplo de consulta respondida de maneira insatisfatória por um sistema de RI convencional. 6
1.6 Exemplo de pré-processamento de consulta ambígua. 7
1.7 Exemplo de consulta respondida com aspectos semânticos. 8

2.1 Conceito como resumo de características. 12
2.2 Exemplo de estruturação de informação em tesauro para o conceito “Bispo”. 15
2.3 Exemplo de estruturação de um mapa de tópicos. 18
2.4 Exemplo de relação semântica do tipo hierárquica. 19
2.5 Arquitetura de funcionamento do algoritmo ESA 29
2.6 Exemplo de texto relacionado com artigos da Wikipédia pelo algoritmo ESA 30
2.7 Exemplo de termo relacionado com artigos da Wikipédia pelo algoritmo ESA 30
2.8 Visão Geral do Projeto RISO-T. 31
2.9 Criação de vetores temáticos de domínio baseado nas informações da Wikipédia. 32

4.1 Exemplo da estruturação da instância *“luva”* no projeto RISO-ES. 45
4.2 Estrutura conceitual do sistema RISO-ES e instância exemplificadora. . . . 46
4.3 Fluxo das etapas para indexação semântica de conceitos. 46
4.4 Estruturação de informações em ontologias. 50
4.5 Modelo de avaliação proposto por este trabalho. 51

4.6 Processo de extração de conceitos de documentos textuais. 56
4.7 Processo de enriquecimento conceitual com relacionamento perfeito. 57
4.8 Processo de enriquecimento conceitual com relacionamento parcial. 58
4.9 Processamento de triplas em momento inicial. 60
4.10 Processamento de triplas envolvendo afirmações em que não é possível estabelecer algumas relações. 60
4.11 Enriquecimento conceitual com os relacionamentos presente nas fontes externas de informação. 61
4.12 Estruturação de informações em ontologias utilizando RDF. 62
4.13 Construção dos conjuntos de elementos de domínios distintos. 66
4.14 Processo de desambiguação implementado pelo RISO-ES. 66
4.15 Grafo minimal e conexão com o documento. 67
4.16 Grafo minimal e estrutura de tópicos semânticos representados computacionalmente. 67

5.1 Exemplo de execução da coleção Reuters-21578. 72
5.2 Cobertura do RISO-T versus UBY para coleção Reuters-21578. 72
5.3 Cobertura RISO-T versus UBY para Analisadores Morfológicos. 74
5.4 Construção dos vetores para o cálculo do Teste-t pareado. 78
5.5 Teste-t para ambas as hipóteses alternativas de igualdade e superioridade. . 78
5.6 Utilização do algoritmo ESA para quantificar o enriquecimento semântico proporcionados pelo UBY e RISO-T. 80
5.7 Melhor enriquecimento ponto a ponto. 80
5.8 Melhor média entre os elementos propostos entre RISO-T e UBY. 81
5.9 Tempo gasto em minutos para processar coleções de Analisadores Morfológicos entre RISO-T e UBY. 83
5.10 Tempo gasto em minutos para processar as subcoleções Reuters-21578 entre RISO-T e UBY. 84
5.11 Tempo de processamento para calcular o enriquecimento semântico entre RISO-T e UBY. 84

A.1 Conceitos selecionados para o experimento. 95

A.2 Parte do processamento dos conceitos selecionados submetidos ao RISO-ES. 95
A.3 Parte do processamento dos conceitos selecionados submetidos ao UBY. . . 96
A.4 Parte da execução dos conceitos selecionados submetidos ao UBY. 97

Lista de Tabelas

2.1 Definições das características de uma ontologia. 26

5.1 Comparação da cobertura entre UBY e RISO-T para a coleção Reuters-21578. 72
5.2 Número médio de conexões entre UBY e do RISO-T para a coleção Reuters-21578. 73
5.3 Comparação da Cobertura entre UBY e RISO-T para Analisadores Morfológicos. 74
5.4 Conectividade do UBY e RISO-T quando submetidos às coleções de Analisadores Morfológicos. 75
5.5 Quadro comparativo da validação da qualidade de enriquecimento semântico entre UBY e RISO-T. 79
5.6 Tempo de processamento entre UBY e RISO-T quando submetidos aos Analisadores Morfológicos. 82
5.7 Tempo de processamento entre UBY e RISO-T quando submetidos às subcoleções Reuters-21578. 83

Lista de Códigos Fonte

4.1 Construção de subgrafos 64

Capítulo 1

Introdução

O processo de globalização, acelerado por meio da internet, permitiu ao ser humano potencializar seus processos de comunicação, contribuindo significativamente para a intensificação da produção de conhecimento e, consequentemente, facilitando a aquisição de informação. Nesse contexto, o usuário comum de internet, passou a utilizar este canal para publicar informações e, assim, gradativamente foi deixando de ser um mero leitor de conteúdo e se transformando em produtor de conhecimento.

Diante do grande volume de informação digital acessível nos últimos anos, a qualidade dos sistemas de Recuperação de Informação (RI) tornou-se essencial para que os usuários destes sistemas pudessem economizar recursos (e.g., tempo, energia) na atividade de busca por conteúdo útil. Com essa finalidade, muitos engenhos de busca atuais, tais como, Google, Yahoo, Bing, entre outros, foram desenvolvidos. Entretanto, não é raro observar usuários demorando minutos durante o processo de busca, avançando páginas sucessivas de resultado e não encontrando nenhuma informação relevante. Perguntas recorrentes como: *"Deverei modificar a consulta para melhorar a qualidade dos resultados apresentados?"* ou *"Deverei seguir buscando as páginas seguintes do resultado?"* refletem a fragilidade da indexação convencional diante do volume de informação e, assim sendo, distancia estas ferramentas de seu real propósito: proporcionar um conjunto resposta que seja útil.

Basicamente, o processo de recuperação de informação começa quando o usuário emite uma solicitação de consulta, geralmente expressa em linguagem natural. Tal solicitação é processada por um engenho de busca que tenta realizar uma correspondência entre as informações presentes na consulta e os índices criados para apontar para os documentos corres-

pondentes [3]. Estes índices, constituem a ponte entre a consulta emitida pelo usuário e os objetos que elas representam, sejam estes: *links*, documentos, vídeos, imagens, sons, mapas, entre outros [38].

A complexidade inerente à linguagem natural, permite a transmissão de uma mesma ideia de inúmeras maneiras distintas, como por exemplo: *"mandioca"*, *"aipim"*, *"macaxeira"*, *"castelinha"*, *"maniva"*, *"maniveira"*, *"pão-de-pobre"* são nomes populares para a espécie de planta *"Manihot Esculenta"*. Cada um desses termos, constituem na verdade, diferentes representações de um mesmo conceito[1]. Sistemas convencionais de RI, indexam documentos por meio de termos e não conceitos, ou seja, elementos são diferenciados apenas por seu aspecto léxico-sintático e, assim sendo, é provável que uma consulta que contenha, por exemplo, o termo *"mandioca"*, retorne um resultado satisfatório, enquanto que uma consulta pelo termo *"aipim"* ou *"maniveira"* não encontre nenhum resultado relevante.

É possível perceber por meio do exemplo apresentado na Figura 1.1, que o processo de RI convencional não é suficiente para relacionar a consulta *"cars"* com os índices *"auto"* e *"machine"*. Isso porque sistemas convencionais de RI irão comparar a sequência de caracteres presentes na consulta do usuário com a sequência de caracteres de cada um dos índices da base de dados, retornando algum documento apenas em caso de existir uma correspondência exata entre ambos (consulta do usuário e algum índice da base). Para o exemplo da Figura 1.1, seja T o conjunto de termos (ou *tokens*) do índice, então:

$$cars \notin T$$

Sendo assim, não será possível recuperar nenhum documento. Todavia, documentos que foram indexados com índices: *"auto"* e *"machine"*, por serem sinônimos diretos de *"cars"*, poderiam apresentar informações interessantes para o usuário, pois, conceitualmente, tratam do mesmo conteúdo. Além disso, o usuário comum de sistemas de RI desconhece os índices utilizados para indexar as informações de seu interesse e, assim sendo, frequentemente precisa reajustar as consultas emitidas para melhorar a qualidade dos resultados obtidos, limitando assim, a eficácia desses sistemas.

Seria impraticável para sistemas convencionais de RI, mapear e indexar manualmente todas as possibilidades de representação de um determinado termo. Por isso, quando não

[1] A seção 2.1 apresenta uma definição pragmática de conceito.

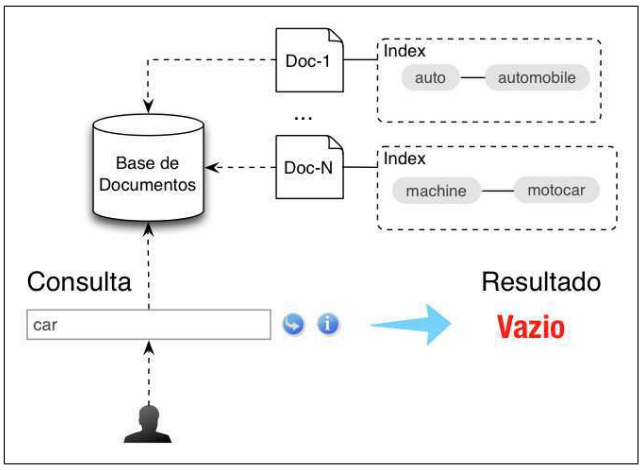


Figura 1.1: Processo de recuperação de informação não semantizado.

há correspondência entre os termos da consulta e os termos indexados, sistemas convencionais tentam abordagens, tais como, sugerir outra consulta ao usuário (“Você quis dizer . . . ?”) como uma forma de evitar a superabundância de resultados inúteis, ou até mesmo ausência de resultados.

Por outro lado, implementando uma abordagem semântica, seria possível a um sistema de RI, processar de maneira mais adequada o que foi escrito pelo usuário e relacionar semanticamente as informações presentes na consulta com as informações previamente utilizadas no processo de indexação. A Figura 1.2 apresenta um exemplo do processo de recuperação de informação utilizando uma abordagem semântica. É possível perceber que, ainda que não existam documentos indexados por *“cars”*, o processo de enriquecimento semântico da consulta é capaz de estabelecer relações semânticas com *“auto”*, *“automobile”* e *“motocar”*. E assim, recuperar documentos que foram indexados por tais índices, que, mesmo não estando indexados diretamente por *“cars”*, seus índices representam a mesma ideia ou conceito e, provavelmente, constituem conteúdo útil. Tal abordagem, favorece a recuperação de informação por ideia ou conceito relacionando semanticamente tanto os termos da consulta com os índices utilizados para representar documentos quanto os próprios índices entre si.

A indexação conceitual ou semântica de documentos consiste em relacionar um conceito extraído de um documento e presente em uma base de conhecimento, com outro conceito por meio de uma relação semântica. Tais relações semânticas são propriedades linguísticas do idioma em questão, podendo ser inferidas automaticamente a partir de fontes de informação

previamente construídas pelo homem como, por exemplo, o WordNet.

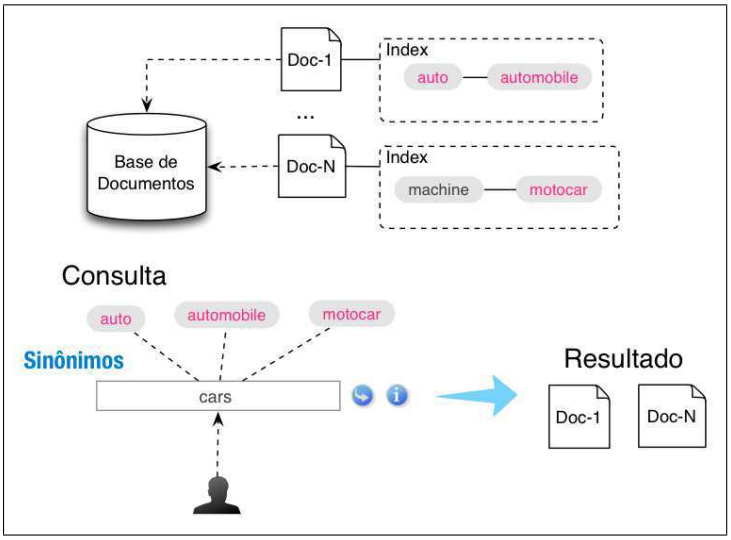


Figura 1.2: Processo de recuperação de informação semantizado.

Uma outra situação em que os sistemas de RI convencionais não são capazes de responder com qualidade satisfatória, é a busca por conjuntos de elementos expressos por meio de um único conceito. Consultas, tais como: *"obras de monet e renoir"*, *"carros luxuosos"*, *"atores da tv americana que nasceram antes de 1970"*, dentre outras, são instâncias de consultas em que a simples análise sintática dos termos da consulta, sem a análise conceitual de seu objetivo, não apresentará resultados satisfatórios. Para um usuário que deseja recuperar todos os documentos de um engenho de busca que façam referência a algum animal (e.g., gato, cachorro, vaca e boi), por exemplo, ele terá uma árdua tarefa. Sendo o sistema puramente sintático, caso o usuário emita a consulta: *"Animais"*, o sistema irá buscar por informações indexadas por *"Animais"*, em lugar de processar conceitualmente o objetivo da consulta, como ilustra a Figura 1.3.

Para obter o que necessita utilizando um sistema de RI convencional, tal usuário terá que criar uma lista contendo todos os animais e, para cada elemento da lista, realizar uma consulta no sistema guardando posteriormente os resultados obtidos. Por outro lado, em um sistema semântico, a consulta por *"Animais"*, já seria suficiente para que o sistema relacionasse as informações que este conceito compreende e fornecesse satisfatoriamente as informações pretendidas como evidenciado pela Figura 1.4.

Fenômenos linguísticos como: polissemia, ambiguidade, sinonímia e acronímia não con-

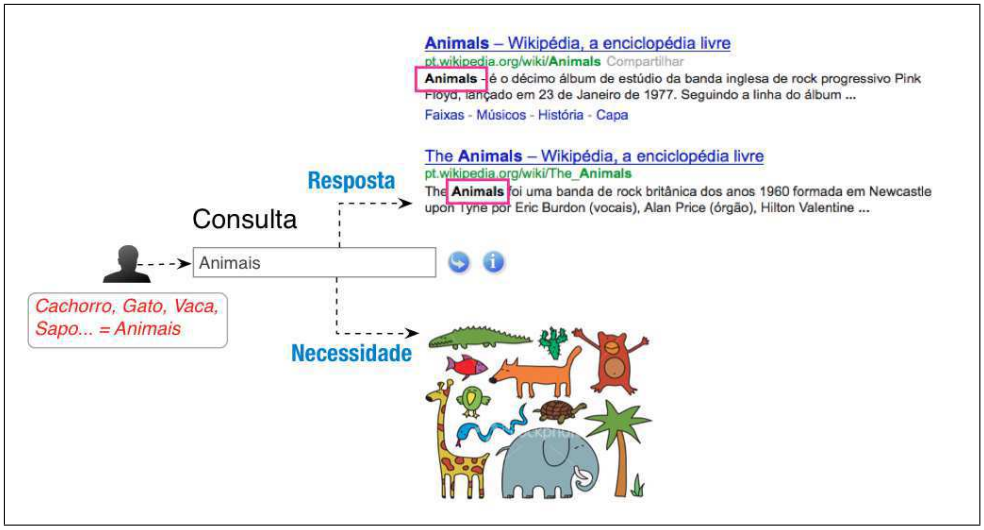


Figura 1.3: Consulta conceitual com interpretação sintática.

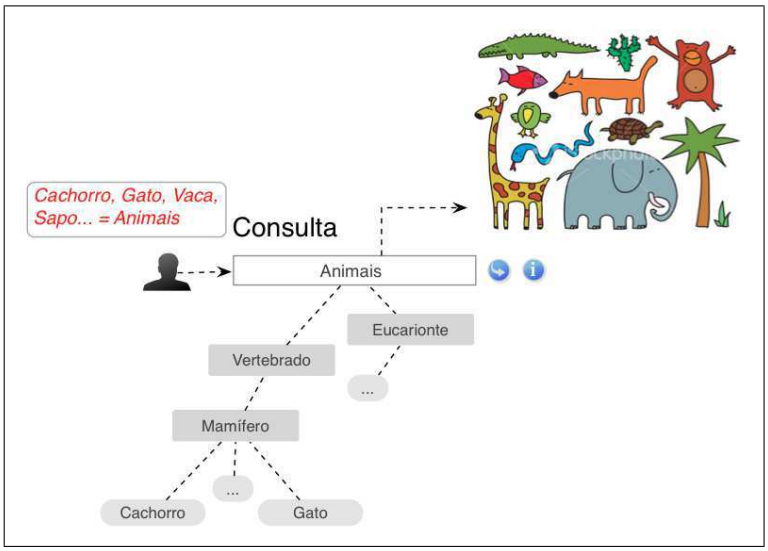


Figura 1.4: Consulta conceitual com interpretação semântica.

seguem esclarecer de maneira objetiva a ideia do usuário no momento de emitir sua consulta. Por exemplo, para a consulta *"informações sobre mangas"*, tem-se um caso típico de polissemia. O sistema convencional não saberá distinguir entre os mais diversos conceitos: *"parte do vestuário que cobre o braço"*, *"fruta"*, *"chocalho grande"* e a *"ação de zombar fingindo seriedade"*. Um outro exemplo que reflete claramente a complexidade linguística, pode ser percebido por meio de uma consulta pelo acrônimo *"OCL"*. Para este caso, o sistema também não saberá identificar se a intenção do usuário é: *"Object Constraint Language"*, *"Ocean Climate Laboratory"* ou *"Organisation Communiste Libertaire"*, e por isso, inúmeros itens retornados podem ser absolutamente inúteis.

Os motivos expostos ressaltam a importância da construção de métodos, que permitam a compreensão da intenção do usuário ao emitir determinada consulta e da indexação semântica no processo de filtrar informação relevante para usuários de sistemas de RI. A Figura 1.5 exemplifica a experiência de um determinado usuário que busca por informações sobre o acrônimo *"SBC"* no contexto da Ciência da Computação em um sistema de RI convencional. Percebe-se por meio da primeira página de resultados do buscador (que, por meio do algoritmo de *ranking* apresenta os resultados mais "relevantes"), que apenas 1 único elemento é efetivamente útil.

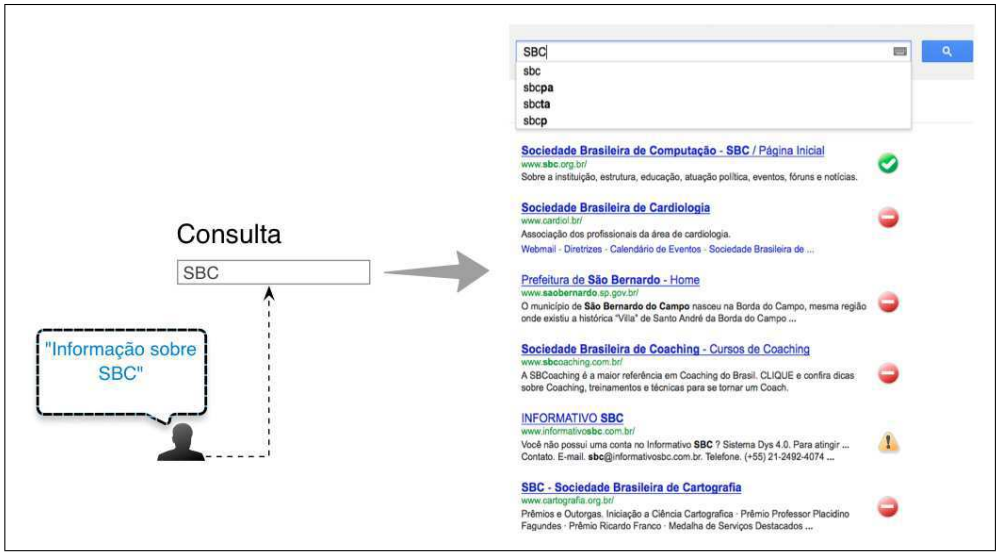


Figura 1.5: Exemplo de consulta respondida de maneira insatisfatória por um sistema de RI convencional.

Sendo o sistema de RI ideal aquele em que todos os documentos recuperados são relevan-

tes para uma determinada consulta realizada pelo usuário, de forma que nenhum documento relevante esteja faltando [28], seria interessante uma abordagem de software que pudesse compreender a intenção do usuário, para então, sugerir-lhe possibilidades de relacionar semanticamente sua consulta com outros elementos de seu interesse e, assim, de acordo com sua escolha, proporcionar a recuperação de uma maior quantidade de informação útil.

Para o caso de uma consulta como *"SBC"*, o fato de permitir a desambiguação por parte do usuário, aumentaria consideravelmente a quantidade de conteúdo útil recuperado. Como evidenciado por meio da Figura 1.6, após ser emitida uma consulta buscando pelo acrônimo *"SBC"*, é realizado um pré-processamento no qual permite-se ao usuário a possibilidade de informar qual(is) o(s) sentido(s) de *"SBC"* que é(são) de seu interesse. No caso do exemplo em questão, após eleger o item relacionado com *"Ciência da Computação"*, informações referentes a *"São Bernardo do Campo"*, *"Sociedade Brasileira de Citopatologia"*, entre outros, não estarão no resultado final e, assim, existirá uma considerável melhoria na precisão dos resultados apresentados.

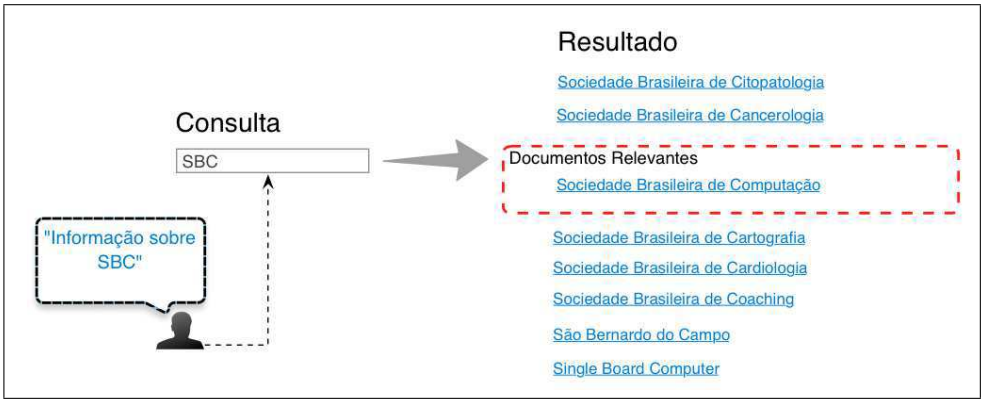


Figura 1.6: Exemplo de pré-processamento de consulta ambígua.

Propõe-se, neste trabalho, um novo paradigma com ênfase na utilização de fontes heterogêneas de informação com características de dicionários, enciclopédias e de sentido comum, com a finalidade de realizar o enriquecimento semântico de informações presentes em documentos. Esta atividade é responsável por obter das fontes heterogêneas um conjunto de termos que esteja semanticamente relacionado com os termos presentes nos documentos, relacionando-os e utilizando estes novos termos para favorecer a recuperação dos próprios documentos. Com isso, melhora-se a cobertura e precisão de sistemas de RI, mediante a

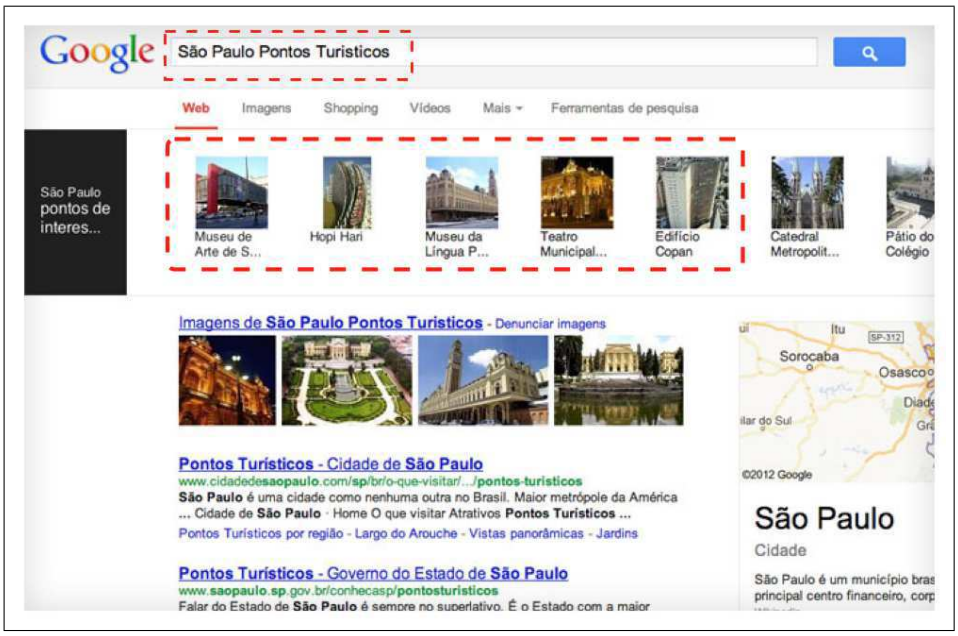


Figura 1.7: Exemplo de consulta respondida com aspectos semânticos.

indexação semântica de termos, em oposição à indexação apenas por palavra-chave. Desse modo, isola-se a ideia das coisas de suas inúmeras formas de representação e se utiliza deste átomo semântico para indexar informação dos mais variados formatos.

Nesta abordagem,*"mandioca"*, *"aipim"*, *"castelinha"*, *"maniva"*, *"maniveira"*, *"pão-de-pobre"* e até mesmo *"Manihot Esculenta"*, apesar de serem diferentes formas de representação de uma mesma ideia, para o paradigma proposto, passam a representar um único conceito. Além disso, se um termo homônimo, como *"manga"*, for desambiguado, conceitos relacionados, como por exemplo, *"fruta"* e *"caroço"*, poderão ser utilizados para enriquecer a recuperação da informação, fornecendo maior dinamicidade entre o usuário e o sistema de RI.

A própria Google, maior empresa no seguimento de RI da atualidade, consciente da necessidade de melhorar a qualidade de seus serviços, lançou em maio de 2012 o Knowledge Graph[2], que vem a ser um incremento semântico de seu buscador convencional, visando suprir algumas das deficiências supracitadas. Na Figura 1.7 é possível perceber que para a consulta *"São Paulo Pontos Turísticos"*, o sistema não se limita apenas em retornar informações indexadas por *"São Paulo"*, *"Pontos Turísticos"*, *"Pontos"* e *"Turísticos"*, mas tenta

[2]http://www.google.com/insidesearch/features/search/knowledge.html

compreender a necessidade do usuário (*"Quero encontrar informações sobre pontos turísticos em São Paulo"*) e relacionar semanticamente a consulta com inúmeros elementos tais como: *"Hopi Hari"*, *"Museu da Língua Portuguesa"*, que possuem grafia absolutamente distinta, entretanto, forte relação semântica com a consulta emitida.

Apesar do avanço, o sistema ainda apresenta fragilidades, principalmente relativas ao grau de desambiguação necessário para a efetiva consolidação do aspecto semântico e detecção exata do conceito. Ainda que a informação *"Pontos Turísticos"* tenha evitado que *"São Paulo"* seja interpretada como: *"santo"*, *"time de futebol"*, entre outras interpretações, percebe-se, pelo exemplo, que ainda não foi possível saber a qual *"São Paulo"* se refere a consulta, se ao *estado de São Paulo*, ou se a *capital do estado de São Paulo* que possuem a mesma grafia e classificação como região geográfica. Geralmente, debilidades e incoerências do gênero são causados pela incipiência da base de conhecimento, que são pouco abrangentes a ponto de cobrir toda imensidão de detalhes de interpretações do mundo real.

Desta forma, deve-se considerar que, para o funcionamento adequado da RI semântica, se faz indispensável a implementação do processo de enriquecimento semântico tanto durante a indexação de documentos textuais quanto durante a interpretação dos elementos da consulta do usuário. Desta maneira, será possível estabelecer um relacionamento entre os termos da consulta e os termos utilizados na etapa de indexação por meio do conteúdo da informação e não apenas por meio dos vocábulos que cada módulo (consulta e indexação) contém (como é feito na RI convencional).

1.1 Objetivos

A seguir, apresentam-se os objetivos geral e específico propostos neste trabalho.

1.1.1 Objetivo Geral

Propor uma nova abordagem capaz de permitir a indexação e o enriquecimento semântico de conceitos presentes em documentos textuais.

1.1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

1. Propor um algoritmo que seja capaz de obter de diferentes fontes externas e heterogêneas:(*WordNet, Wikipédia, Wiktionary, DBPedia, Conceptnet5, Verbosity e Reverb*) termos semanticamente relacionados aos termos presentes nos documentos relacionando-os e armazenando-os em um único ambiente.

2. Relacionar as informações semânticas obtidas do item 1 com os conceitos previamente marcados em documentos textuais.

3. Adaptar algoritmos que possibilitem a desambiguação de termos oriundos do enriquecimento semântico para a correta associação com os conceitos presentes no documento.

4. Indexar semanticamente documentos textuais de uma base de dados e estruturar as informações dos itens 1, 2, 3 e 4 em mapa de tópicos[3].

5. Avaliar a abordagem proposta, comparando-a com um projeto de finalidade semelhante.

1.2 Relevância

A indexação semântica de documentos é algo potencialmente útil, uma vez que relaciona documentos por meio de seu conteúdo e não apenas pelos vocábulos que contêm. Tal processo, permite ao usuário desses sistemas recuperar documentos conceitualmente relacionados às necessidades de informação, ou seja, além de possibilitar a recuperação de documentos que contemplem os termos fornecidos em uma consulta, tal abordagem possibilita recuperar documentos que possuem o mesmo sentido ou ideia dos elementos da consulta, em oposição à sistemas convencionais que se limitam a realizar o cálculo da distância vetorial entre vetores (*bag of words*) construídos com os termos da consulta e dos documentos presentes na base de dados, sem tentar compreender seu propósito ou sentido.

[3]Definição e explicação sobre a estrutura de mapa de tópicos encontram-se na seção 2.2.5

Para a Web Semântica, subárea da Ciência da Computação que busca atribuir um significado (sentido) aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo humano como pelo computador, este trabalho de investigação contribui concebendo uma novo paradigma capaz de relacionar conhecimento de distintas fontes de informação, tais como, dicionários, enciclopédias e sentido comum unindo-as em um único ambiente. Ampliando, dessa maneira, a capacidade das máquinas de melhor processarem a linguagem humana.

Durante a revisão da literatura realizada na etapa de escrita dessa dissertação, não foi encontrado nenhum método ou ferramenta de construção automática de mapa de tópicos que englobasse análises de fontes externas e de conteúdo heterogêneo ao processo de enriquecimento semântico de conceitos aqui propostos, mas sim, inúmeros exemplos de projetos que poderiam se beneficiar deste trabalho, tais como, o tesauro produzido pelo National Cancer Institute (NCI[4]).

Em suma, o desenvolvimento de um sistema computacional com as características apresentadas pela nova abordagem proposta nesse trabalho é útil para auxiliar a atividade de recuperação de informação relevante para o usuário, tornando a linguagem humana mais compreensível para os computadores. Além disso, permite modelar o conhecimento presente em documentos textuais, desempenhando um papel fundamental no contexto do projeto RISO-T, sendo responsável, neste contexto, pelo enriquecimento semântico de documentos, indexação semântica, bem como, construção do mapa de tópicos. A seção 2.8 apresenta detalhes do projeto RISO-T e explica como este módulo está inserido neste contexto.

1.3 Escopo de Trabalho

Este trabalho apresenta-se como um módulo para o sistema de Recuperação Semântica de Objetos Textuais (RISO-T) sendo responsável pela indexação e enriquecimento semântico de conceitos presentes em documentos textuais.

[4]http://ncit.nci.nih.gov/

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta a terminologia e fundamentação teórica utilizadas ao longo deste trabalho de dissertação.

2.1 Termo e Conceito

Segundo Studer et al. [59], um conceito constitui o elemento básico de uma terminologia estabelecida para o domínio de um problema. Para Bui e Duc [6], um conceito é uma representação abstrata e universal das coisas. Para Wuster [41] e Peter Becker [61], um conceito se apresenta como um conjunto de características individuais comuns de um determinado objeto percebidas pelos seres humanos em sua intenção ou simplesmente um elemento do pensamento e, esta noção de conceito é adotada neste trabalho. Em outras palavras, o conceito de um objeto chamado de *"carro"*, por exemplo, é o resumo de características que esse objeto possui. A Figura 2.1 ilustra a definição de conceito proposta por Wuster e aceita por este trabalho.

Figura 2.1: Conceito como resumo de características.

Os termos, por sua vez, são representações simbólicas destes conceitos dentro de uma linguagem, utilizando para isso um conjunto de caracteres. Diferentemente dos conceitos, os termos podem ser ambíguos e, assim, carregar mais de uma interpretação. Isso acontece quando um mesmo termo é capaz de representar mais de um conceito, como por exemplo: *"manga"*(de camisa? fruta?) e *"jaguar"*(carro? animal?). Assim sendo, termos ambíguos necessitam ser contextualizados para descreverem apenas um único conceito. Por outro lado, termos ao invés de ambíguos, podem ser sinônimos, neste caso, distintos termos representam o mesmo conceito, por exemplo: $\{pomme, manzana, apple\}$ são diferentes termos que representam exatamente o mesmo conceito. Termos comumente são utilizados no processo de indexação de documentos, porém, o desafio deste trabalho consistiu de realizar uma indexação conceitual.

2.2 Representações do Conhecimento

A Representação do Conhecimento ou Linguagens Documentárias (LDs) são linguagens construídas de maneira artificial a partir de sistemas simbólicos, que visam descrever sinteticamente conteúdos documentais e são utilizadas nos sistemas informacionais para indexação, armazenamento e recuperação da informação. A seguir é aberta uma discussão sobre algumas destas estruturas que possuem relação com o trabalho apresentado no capítulo 4.

2.2.1 Tesauros

O termo *"tesauro"* tem origem no dicionário analógico de Peter Mark Roget em 1852, intitulado *"Thesaurus of English words and phrases"* [53]. Este trabalho, teve como principal objetivo facilitar as atividades literárias do autor no momento de ponderar palavras para serem utilizadas em sua escrita e, assim, começou a estruturar palavras pelo seu significado, ao invés da ordem alfabética, como ocorria com os dicionários convencionais da língua naquela época. Seu objetivo inicial foi o de agrupar palavras de acordo com as ideias que elas tentam exprimir, ponderando, ainda que de maneira rudimentar, sua intensidade. Dessa maneira, as palavras foram arranjadas estritamente de acordo com seu significado.

Gomes e Campos [8] definem um tesauro como sendo uma linguagem documentária dinâmica que contém termos relacionados semântica e logicamente, cobrindo de modo com-

preensivo um domínio do conhecimento.

Segundo Cavalcanti [9], um tesauro constitui uma lista estruturada de termos associados, empregada por analistas de informação e indexadores, para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação procurada.

O Tesauro, também conhecido como dicionário de ideias afins, é uma lista de palavras com significados semelhantes, dentro de um domínio específico de conhecimento que tem como função principal, o controle terminológico do vocabulário utilizado em uma área específica do conhecimento, indicando as relações entre conceitos [68]. Todo tesauro constitui uma linguagem especializada, estruturada conforme rede conceitual e que apresenta relações semânticas entre os termos tanto hierárquicas (gênero/espécie; todo/parte) quanto associativas. Por outro lado, tesauros não incluem definições acerca de vocábulos, pelo menos não detalhadas, uma vez que essa atividade é de competência dos dicionários [56]. OAD Thesaurus do National Institute on Alcohol Abuse and Alcoholism (NIAAA[1]), APAIS[2], Government of Canada Core Subject Thesaurus[3] e Library of Congress Thesauri[4] são exemplos de projetos atuais que desenvolveram tesauros para facilitar a aquisição e recuperação de informação.

2.2.2 Dicionários

Um dicionário é uma compilação de palavras ou de termos próprios ou, ainda, de vocábulos de uma língua, que em linhas gerais, busca explicar o sentido de um determinado item léxico individual valendo-se de outros itens léxicos [2].

Quase sempre dispostos em ordem alfabética, cada dicionário possui classificações em harmonia com objetivos e finalidades didáticas aos quais se compromete em abranger. Isso se deve a uma constante necessidade de atender aos diversificados níveis e áreas de conhecimento.

Quanto aos tipos, comumente encontram-se:

- *Dicionários gerais da língua:* são de versão extensa ou com adaptação a usos escola-

[1] http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

[2] http://www.nla.gov.au/apais/thesaurus/index.html

[3] http://www.thesaurus.gc.ca/recherche-search/thes-eng.html

[4] http://www.loc.gov/lexico/servlet/lexico/

res. Possuem um considerável número de palavras, definidas em suas várias acepções e significados. Exemplos: *Dicionário Michaelis*, *Dicionário Aurélio*, entre outros.

- *Dicionários etimológicos:* fornecem a origem de cada palavra por meio de sua formação e evolução.
- *Dicionários temáticos:* organizam vocabulários específicos de determinada ciência, arte ou atividade técnica. Como, por exemplos: *Dicionário Jurídico*, *Dicionário de Comunicação*, de *Astronomia* e *Astronáutica*.
- *Dicionários de abreviaturas:* úteis por facilitarem a comunicação ainda mais nesta época de abreviaturas e siglas.
- *Dicionários bilíngues ou plurilíngues:* explicam o significado dos vocábulos estrangeiros e suas correlações com os vocábulos nativos.

Além dos dicionários supracitados, ainda existem outros que se propõem a atender diversas finalidades como dúvidas e dificuldades de uma língua, frases feitas, provérbios, gírias e expressões regionais, entre outros.

Comparando a representação por tesauro, com a representação de dicionários, sob um ponto de vista prático, é possível perceber que existe diferença entre as representações. Por exemplo, enquanto o modelo de *dicionário* representa o item léxico "Bispo" como sendo:

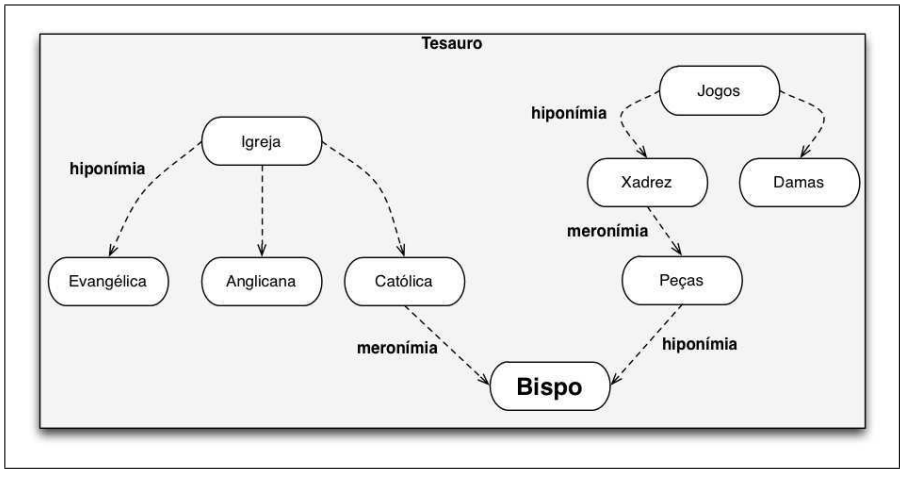


Figura 2.2: Exemplo de estruturação de informação em tesauro para o conceito "Bispo".

- *"Padre que na Igreja Católica recebe, através da sagração, a plenitude do sacerdócio e que tem a direção espiritual de uma diocese. Em muitas Igrejas protestantes, dignidade eclesiástica."*

- *"Peça de jogo de xadrez."*

Por sua vez, em um modelo baseado em *tesauro*, uma possível representação para o conceito pode ser encontrada na Figura 2.2.

2.2.3 Enciclopédia

Uma enciclopédia constitui uma coletânea de textos, geralmente ordenada lexicograficamente, cujo objetivo principal é descrever da melhor maneira possível o estado atual do conhecimento humano, assim sendo, pode-se defini-la ainda, como uma obra que trata de todas as ciências e artes do conhecimento do homem atual, podendo ser tanto um livro físico de referência para praticamente qualquer assunto do domínio humano, como também uma obra na internet [27].

2.2.4 Rede Semântica

Uma rede semântica é uma notação gráfica para representar o conhecimento em padrões de nós interconectados a arcos. Nesta estrutura, nós são conceitos; enquanto os arcos representam as relações semânticas utilizadas para relacionar cada um desses conceitos. Implementações computacionais de redes semânticas foram inicialmente desenvolvidas para a inteligência artificial, entretanto, muito antes disso, o termo foi utilizado em filosofia, psicologia e linguística [37; 19; 33; 70].

O que é comum a todas as redes semânticas é uma representação declarativa gráfica, que pode ser usada tanto para representar o conhecimento, como para apoiar sistemas automatizados para o raciocínio sobre o conhecimento. Apesar de algumas versões serem informais, a grande maioria é definida por lógicas formais.

A seguir, definem-se os seis tipos mais comuns de redes semânticas:

- *Redes de definição:* enfatizam subtipos e a relação *É-um*, também chamado de hierarquia de generalização ou subsunção, suporta as regras de herança para obter proprieda-

des definidas de um super-tipo para todos os seus sub-tipos. As afirmações presentes nessas redes são necessariamente verdade.

- *Redes de Afirmações:* são projetadas para afirmar proposições. Ao contrário das redes de definição, a informação de uma rede de afirmações é assumida como sendo contingentemente verdadeiro, a menos que seja explicitamente marcado com um operador modal. Algumas redes de afirmações têm sido propostas como modelos de estruturas conceituais subjacentes à semântica da linguagem natural.

- *Redes de implicações:* usam implicação como a relação primária para ligar os nós. Eles podem ser usados para representar padrões de crenças, causalidade ou inferências.

- *Redes Executáveis:* incluem mecanismos ou procedimentos anexados capazes de realizar inferências, mensagens secretas, ou busca por padrões e associações.

- *Redes de Aprendizagem:* constroem ou ampliam suas representações por meio da aquisição de conhecimento a partir de exemplos. O novo conhecimento pode mudar a antiga rede pela adição ou exclusão de nós e arcos, ou modificando valores numéricos de pesos associados aos nós e arcos.

- *Redes Híbridas:* Redes híbridas combinam duas ou mais das técnicas anteriores.

2.2.5 Mapa de Tópicos

Mapas de tópicos são estruturas utilizadas para representar o conhecimento compostas por 3 elementos: *tópicos, associações e ocorrências*. Um tópico consiste de um nó, que representa um conceito ou elemento que, por sua vez, relaciona-se com outros tópicos por meio de *associações* que são relações de tópicos de tipo n-ary[5]. As características de um *tópico* são contextualizadas por meio do elemento *escopo*, que consiste, em linhas gerais, de uma descrição com objetivo de tornar o conceito único [25].

Basicamente, os *mapas de tópicos* consistem de uma extensão do modelo de rede semântica, acrescentando-lhe, além das relações entre tópicos (conceitos), associações com *ocorrências* que indicam o documento em que determinado conceito é encontrado, ou seja,

[5]http://www.w3.org/TR/swbp-n-aryRelations/

de maneira genérica, podem ser, imagens, vídeos, sons ou, em nosso caso, documentos textuais. Além disso, os mapas de tópicos são padronizados por meio da norma ISO 13250, diferentemente das redes semânticas, que não possuem padronização.

O diagrama da Figura 2.3 apresenta um exemplo da estrutura de mapa de tópicos[6].

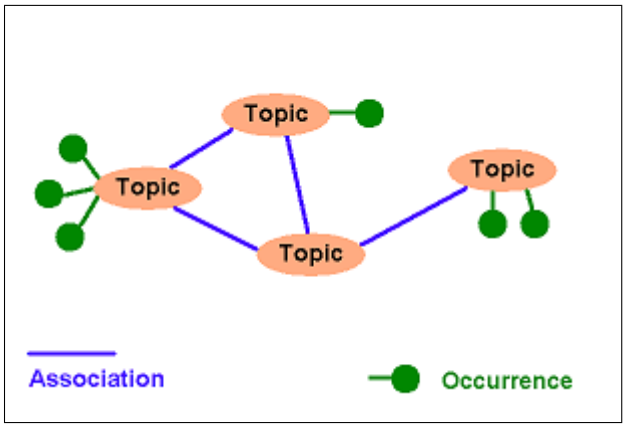


Figura 2.3: Exemplo de estruturação de um mapa de tópicos.

O mapa de tópicos, representa os principais conceitos descritos tanto em bases de dados quanto em documentos, relacionando-os. Assim, quando um documento contém a seguinte afirmação: *"O procedimento de manutenção de X consiste nos seguintes passos . . . "* o mapa de tópicos é capaz de relacionar a informação deste documento com a informação presente em outros documentos e ampliar esta informação de *"X"* para, por exemplo: *"Parte X é do tipo Q e está contida em partes Y e Z e seu procedimento de manutenção reside no documento W"*. Isto significa gerir o significado da informação e seus relacionamentos, ao invés de apenas a informação [24].

Neste trabalho, utiliza-se esta riqueza estrutural relacionando conceitos (tópicos) como uma estratégia para suprir a necessidade de indexação léxico-semântica convencional.

Na construção de um mapa de tópicos, por exemplo, é estabelecido um controle do vocabulário visando que cada conceito seja expresso por um único e inequívoco termo ou descritor. Para este fim, utilizam-se de várias fontes, tais como tesauros da mesma área ou área afim, dicionários, vocabulários, esquemas de classificação, índices de publicações periódicas, assim como outros documentos da literatura especializada em que se irão controlar os termos. [42; 66; 57; 43]

[6]Esta imagem foi retirada de http://en.wikipedia.org/wiki/TopicMaps/.

2.3 Relações Semânticas

A semântica é a parte da gramática que estuda aspectos relacionados com o sentido das palavras e suas possibilidades de interpretação dentro de uma língua. Nesse sentido, as *relações semânticas* são as relações de significado existentes entre conceitos, isto é, a correta compreensão e associação entre o sentido e as ideias representadas por uma palavra e por outras com as quais estabelece relação [41], não importando sua grafia.

Em [16], Dahlberg estabelece que estas relações podem ser classificadas como:

- *Relações Hierárquicas ou Abstrativas:* Quando as características apresentadas por um conceito englobam as características apresentadas por outro conceito, estabelecendo-se assim, uma relação de generalidade/especificidade. A Figura 2.4 exemplifica uma relação hierárquica entre árvore e macieira, isso porque, toda macieira, necessariamente é uma árvore. Outra relação hierárquica é a relação *instância-de* que relaciona instâncias e classes.

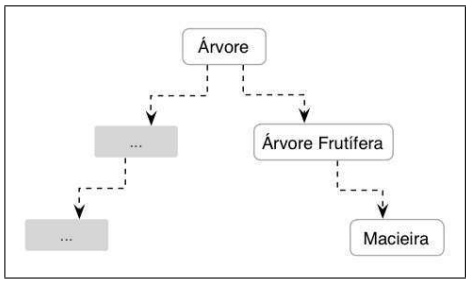


Figura 2.4: Exemplo de relação semântica do tipo hierárquica.

- *Relações Partitivas:* Quando um dos conceitos é parte do outro, ou seja, relação todo / parte. Exemplo:

$$\text{Parte-De}(\text{Árvore}, \textit{\{raízes, tronco, galhos, folhas, flores, frutos\}})$$

- *Relações de Oposição:* Quando um conceito é oposto, ou seja, quando expressa uma ideia contrária a outro com quem se compara. Exemplo:

$$\text{Antônimo-De}(preto, branco)$$

Tanto as *relações hierárquicas* como as *partitivas* referem-se principalmente às relações entre os objetos, enquanto as *relações de oposição* referem-se mais à relação entre propriedades de conceitos.

- *Relações Funcionais:* Aplicam-se a conceitos que expressam processos. Exemplo:

  Relação-Funcional-Com(Produção, *{produto,produtor,comprador}*)

- *Relações Associativas:* Relações que ocorrem entre conceitos no cotidiano. Exemplo:
  - Automação - Computadores.
  - Ação - Resultado.
  - Tecelagem - Tecidos.

Weiszflog e Lyons [67; 44] respectivamente, em seus trabalhos, apresentam várias relações semânticas, dentre elas:

- *Sinonímia:* É a relação que se estabelece entre dois ou mais conceitos que apresentam significados iguais ou semelhantes, isto é, sinônimos. Exemplos:
  - Cômico - engraçado.
  - Débil - fraco, frágil.
  - Distante - afastado, remoto.
- *Antonímia:* É a relação que se estabelece entre dois ou mais conceitos que apresentam significados diferentes, contrários, isto é, antônimos. Exemplos:
  - Economizar - gastar.
  - Bem - mal.
  - Bom - ruim.
- *Homonímia:* É a relação entre dois ou mais conceitos que apesar de possuírem significados diferentes, possuem a mesma estrutura fonológica.

  As homônimas podem ser:

  - *Homógrafas:* palavras iguais na escrita e diferentes na pronúncia. Exemplos:
    - gosto (substantivo) - gosto / (conjugação verbo gostar).
    - conserto (substantivo) - conserto (verbo consertar).
  - *Homófonas:* palavras iguais na pronúncia e diferentes na escrita. Exemplos:
    - cela (substantivo) - sela (verbo).
    - cessão (substantivo) - sessão (substantivo).
    - cerrar (verbo) - serrar (verbo).
  - *Perfeitas:* palavras iguais na pronúncia e na escrita. Exemplos:
    - cura (verbo) - cura (substantivo).
    - verão (verbo) - verão (substantivo).
    - cedo (verbo) - cedo (advérbio).

- *Paronímia:* É a relação que se estabelece entre duas ou mais palavras que possuem significados diferentes, mas são muito parecidas na pronúncia e na escrita, isto é, os parônimos. Exemplos:
  - cavaleiro - cavalheiro.
  - absolver - absorver.
  - comprimento - cumprimento.
  - aura (atmosfera) - áurea (dourada).
  - conjectura (suposição) - conjuntura (situação decorrente dos acontecimentos).
  - descriminar (desculpabilizar) - discriminar (diferenciar).

- *Polissemia:* É a propriedade que uma mesma palavra tem de apresentar vários significados. Exemplos:
  - jaguar (carro ou animal ?).
  - manga (vestuário ou fruta ?).

- *Meronímia:* O conceito que denota uma parte relativamente a um todo (holônimo). Exemplo:

  - *Manga e punho* são merônimos de *camisa*.

- *Hiperonímia:* Relação semântica de ordem hierárquica que um conceito assume em relação a outro (o hipônimo) em virtude da sua maior abrangência de sentido. O hiperônimo é etimologicamente um nome que está numa posição hierárquica superior (hiper) por ser capaz de incluir outras palavras - os seus hipônimos.Exemplo:

  - *Animal* é um hiperônimo de *cão, gato, leão, tigre, elefante e girafa*.

- *Hiponímia:* Relação semântica em que uma palavra está num plano hierárquico inferior, uma vez que pertence a uma classe ou espécie que a inclui ao nível do significado. Este fato implica que o significado do hipônimo (etimologicamente significa nome pequeno) é mais específico e mais restrito do que o significado do hiperônimo a que pertence. O conceito de hiponímia também só é entendido em relação ao conceito de hiperonímia. Exemplo:

  - *sardinha, salmão, pescada e bacalhau* são hipônimos de *peixe*.

2.4 Fontes Externas e Heterogêneas de Informação

Fontes externas de informação se referem a todo conteúdo que não é gerado ou mantido pelo RISO-T ou não é proveniente diretamente dos documentos ou da consulta do usuário. Tais fontes englobam, por exemplo: *WordNet* [21; 63], *Wikipédia* [65], *Reverb* [20], *Verbosity* [49] e *Wiktionary* [71].

Muitos recursos presentes na web - como, por exemplo, a Wikipédia[7] e Wiktionary[8] - utilizam-se da chamada inteligência coletiva para disponibilizar informação precisa sobre os mais variados temas [27]. Outros recursos, como, por exemplo, o ConceptNet5[9], contêm informações sobre o sentido comum das coisas e provêem informações sobre utilização de instâncias de conceitos.

[7] http://www.wikipedia.org/

[8] http://www.wiktionary.org/

[9] http://conceptnet5.media.mit.edu/

Uma outra vertente de fonte de informação é obtida por meio de ferramentas como a WordNet[10], que apresentam conteúdo léxico-semântico de conceitos, seus significados e relações semânticas entre eles.

Apesar de a obtenção de conceitos e suas informações por meio das fontes de informação descritas serem simples, uní-las em uma única estrutura e empregar o resultado obtido no processo de indexação e recuperação de informação não constitui uma tarefa trivial, justamente pelo fato de não existir nenhuma ponte de comunicação que vincule um determinado conceito presente no WordNet, por exemplo, com outro conceito da Wikipédia.

Na proposta de ambiente de indexação e recuperação semântica de informações, denominado RISO-ES, os conceitos extraídos de documentos textuais são classificados, desambiguados e enriquecidos pelo acesso a estas fontes de informação mencionadas, condensando os diferentes tipos de informação conceitual em uma única estrutura semântica. O trabalho se concentra na determinação dos conceitos que, uma vez obtidos, são enriquecidos e relacionados a outros conceitos, para posteriormente, serem estruturados em mapa de tópicos. Dessa forma, cada conceito, passa a ser um tópico do mapa. As ocorrências da estrutura de mapa de tópicos são preenchidas por documentos que contenham termos que expressem a ideia que o tópico se refere. As associações entre os tópicos se darão por meio de relações semânticas da idioma padrão.

2.4.1 WordNet

Dentro da estrutura do RISO-ES, a WordNet fornece informações conceituais, relações semânticas entre esses conceitos, tais como: sinônimos, hipônimos, hiperônimos, acrônimos, merônimos e polissemia, além de possíveis interpretações para um determinado vocábulo. Com isso, torna-se possível estabelecer as associações entre os tópicos (conceitos), construindo uma estrutura conceitual dentro da estrutura de mapa de tópicos. Da WordNet, por exemplo, é possível inferir que o conceito *"banana"* possui uma relação semântica de hiponímia com *"fruta"*, por se tratar de um termo mais específico, por sua vez, *"fruta"* estabelece com *"banana"* uma relação semântica de hiperonímia, por ser um conceito mais abrangente.

Apesar de descrever inúmeros vocábulos, as informações presentes na WordNet não são completas. Percebeu-se que, para algumas relações semânticas, como é o caso da acronímia,

[10]http://wordnet.princeton.edu/

muitos elementos ainda não foram incorporados. No intuito de suprir esta deficiência, este trabalho incorporou em sua estrutura as informações presentes na Wikipédia, como forma de ampliar a estrutura semântica conceitual para os acrônimos, que são, de maneira pragmática, sinônimos produzidos com as iniciais de cada termo. Além disso, informações de conhecimento de mundo ou enciclopédicas também não são contemplados pelo WordNet. Por exemplo, *"big"* para o WordNet é um adjetivo qualificativo, sinônimo de *"large"*. Entretanto, como conceito enciclopédico, *"big"* não é apenas isso. Para milhares de pessoas o conceito representado por *"big"* refere-se, também, ao filme estrelado em 1988 por Tom Hanks e escrito por Gary Ross. Tais informações estão presentes na Wikipédia. A Figura 4.2 apresentou uma instância de conceito para o vocábulo *"big"* e como seria sua representação no arcabouço arquitetural proposto neste trabalho.

2.4.2 Reverb

Reverb[11] é um sistema de software que automaticamente identifica e extrai relacionamentos binários de sentenças em inglês. Foi projetado, inicialmente, para ser utilizado na Web como extrator de informações, pois, neste ambiente, as relações não estão explícitas. Apesar de todo esse potencial, o RISO-ES ainda não explora essa funcionalidade para processar seus documentos textuais.

Utiliza-se do projeto Reverb, apenas a base de dados disponibilizada de 15 milhões de afirmações, resultantes após o processamento da coleção ClueWeb09[12] criada para apoiar a investigação sobre recuperação de informações e tecnologias relacionadas com a linguagem humana, composta por cerca de 1 bilhão de páginas da web em 10 línguas distintas.

2.4.3 Wikipédia e Wiktionary

A Wikipédia é uma enciclopédia web, multilíngue, construída por meio da "inteligência coletiva", sem fins lucrativos e de propósito geral, onde encontram-se milhares de artigos sobre os mais variados assuntos. Para cada entrada são determinadas categorias das quais a entrada se relaciona semanticamente. Estas informações serão utilizadas posteriormente para

[11] http://reverb.cs.washington.edu/

[12] http://lemurproject.org/clueweb09.php/

identificar conceitos e relacionamentos hierárquicos. O Wiktionary, assim como o WordNet, apresenta características estruturais de tesauros expressando conceitos e relações entre estes.

As informações enciclopédicas presentes na Wikipédia e Wiktionary são fundamentais para construção do mapa de tópicos do RISO-ES. Entretanto, as informações destas fontes são apresentadas por meio de artigos escritos em texto corrido, ou seja, são informações não-estruturadas em seu formato natural. A extração de conteúdo útil de informações não-estruturadas exige o desenvolvimento de algoritmos de alta complexidade. Em meio a esta dificuldade, alguns projetos como, por exemplo, JWLP[13] [22] e DBpedia[14] [47], buscaram organizar o conteúdo da Wikipédia e fornecer acesso a suas informações de maneira estruturada, possibilitando dessa forma, maior facilidade no processamento eficaz desse conteúdo, abrindo as portas para que sistemas convencionais pudessem usufruir da inteligência coletiva da Wikipédia de maneira simples. No RISO-ES, por questões de estruturação das informações em formato ontológico, utilizou-se o projeto DBpedia.

2.4.4 Verbosity

Até este ponto, a estrutura conceitual possui informações de dicionário e enciclopédicas. Sabe-se apenas que *"big"* é um sinônimo de *"large"* e *"grown up"* e que também é o nome de um filme estrelado por Tom Hanks, cuja trama relata a história de um garoto que queria crescer *("to be big")*. Entretanto, tipos de coisas *"big"* são informações que também têm significado do sentido comum. Por exemplo, se o contexto é cidade, o sentindo comum indicará que *"São Paulo"*, *"New York"* são *"big"*, no sentido de tamanho, *"Shaquille O'Neal"* é *"big"*, no sentido personalidade e cientista, *"Gandhi"* e *"Albert Einstein"* também são, respectivamente *"big"*. Todas as informações que necessitam de um julgamento popular e validação do sentido comum são obtidas da fonte Verbosity[15]. Nesta fonte, que funciona como uma espécie de jogo virtual, o objetivo é gerar afirmações baseadas na submissão repetida inúmeras vezes da mesma informação, por exemplo, se a afirmação de que o *Empire State Building* é *"big"* for repetida determinada quantidade de vezes por distintos usuários, existem fortes indícios de que essa afirmação constitua, de fato, o sentido comum. Do ponto

[13] http://www.ukp.tu-darmstadt.de/software/jwpl/

[14] http://dbpedia.org/About

[15] http://www.gwap.com/gwap/gamesPreview/verbosity/

de vista formal, obtêm-se do sentido comum, instâncias do conceito em um determinado contexto.

2.5 Ontologias

O termo *ontologia* foi, inicialmente, utilizado pela filosofia, consistindo de um relato sistemático da existência dos entes. Para Gruber [29], uma ontologia é uma especificação explícita de uma conceitualização, que tem como intenção, estudar as categorias de coisas que podem ou não existir em um domínio e que pode ser especificada como uma coleção de conceitos e suas definições declaradas.

Discussões sobre a definição e conceitualização do termo *ontologia* são apresentados em [30; 11] e mesmo sem um consenso sobre definições, compartilham características comuns e são vistas como a tecnologia de consolidação para construção da Web Semântica.

Ontologias são expressadas por meio de uma linguagem formal e são especificadas por meio de classes, relações, axiomas e instâncias definidas na Tabela 2.1. Pode-se estruturar em mapa de tópicos de informações valendo-se de uma ontologia.

Característica	Breve definição
Classes	Utilizadas para descrever os conceitos de um domínio, permitindo organização de forma lógica e hierárquica [50].
Relações	Representam o tipo de interação entre as classes de um domínio e as propriedades presentes nas classes e indivíduos [35].
Axiomas	Utilizados para modelar regras assumidas como verdadeiras no domínio em questão de modo a associá-los com os indivíduos [64].
Indivíduos	Utilizados para representar elementos específicos, dados que juntamente com a definição de ontologia representam a base de conhecimento [35].

Tabela 2.1: Definições das características de uma ontologia.

2.6 Bibliotecas Digitais

As bibliotecas digitais surgiram como um recurso eletrônico hábil para armazenar e organizar grande quantidade de informação digital, ampliando recursos de seu equivalente físico, permitindo outro nível de acesso a um público maior de usuários e novas oportunidades para intercâmbio global e compreensão e utilização do acervo por meio de consultas, geralmente expressas em conceitos digitados em linguagem natural [62].

Comumente, o processo de armazenamento e organização das informações é precedido por uma etapa de *indexação*, que consiste em extrair de tais documentos termos julgados representativos para serem utilizados posteriormente no momento de recuperá-lo por meio de uma consulta [39].

2.7 ESA

Quão relacionados estão *"osso"* e *"cachorro"*? E o que dizer de *"preparar os manuscritos"* e *"escrever um artigo"*? Raciocinar sobre o parentesco semântico de expressões da linguagem natural é rotineiramente realizado por seres humanos, entretanto, continua sendo um obstáculo intransponível para os computadores. Os seres humanos não julgam parentesco textual apenas ao nível de palavras no texto, se não que em um nível muito mais profundo, manipulando conceitos, unidades básicas de significado, com as quais organizam e compartilham seus conhecimentos. Desta forma, o processo de interpretação de um documento específico leva em consideração um contexto muito maior que apenas o conhecimento, envolvendo nesta atividade questões históricas e experiência.

Neste sentido, o ESA [23] se apresenta como um método capaz de representar o significado de textos em um vasto espaço dimensional de conceitos derivados da Wikipédia, usando técnicas de aprendizagem de máquina para construir um interpretador semântico que mapeia fragmentos da linguagem natural para de uma sequência ordenada de conceitos da Wikipédia. De modo que, o texto passa a ser representado por um vetor de conceitos e, assim, a atividade de comparar dois textos resume-se em utilizar a técnica de cosseno para a comparação vetorial proposta em [72].

O algoritmo ESA utiliza a Wikipédia, porque ela é, atualmente, o maior repositório de

conhecimento na web disponível em dezenas de idiomas, tendo em sua versão em inglês, com mais de 400 milhões de palavras em mais de um milhão de artigos. A abordagem de edição aberta, produz notável qualidade, possuindo precisão tão boa quanto a enciclopédia Britânica segundo Giles em [27].

Dado um texto, o primeiro passo do algoritmo ESA é fragmentá-lo em partes, construindo um vetor de fragmentos do texto. Logo após, cada fragmento é processado usando o esquema TF-IDF [26]. Por fim, o módulo denominado de *intérprete semântico* irá iterar sobre cada fragmento, recuperando entradas da Wikipédia que correspondam aos índices invertidos conforme o método de classificação baseado em centróide proposto em [32], fundindo-os em um vetor ponderado de conceitos, que visa representar o texto dado, como apresentado pela Figura 2.5 obtida de .

2.7.1 Rápida Formalização do Algoritmo ESA

Seja,

$G = \{W_i\}$: o conjunto de fragmentos obtidos de um texto (sintagmas).

$< v_1, \ldots, v_n >$: o vetor TF-IDF calculado para cada fragmento onde v_i é o peso de um determinado fragmento W_i.

$< k_1, \ldots, k_n >$: a entidade que quantifica a força de associação entre W_i e um determinado conceito c_j pertencente a Wikipédia, em que, $c_j \in \{c_1, c_2, c_3, \ldots, c_N\}$ e N é o número de conceitos presentes na Wikipédia.

Então, a interpretação semântica para o vetor V é um vetor de tamanho N no qual o peso de cada c_j é dado por:

$$\sum_{W_i \in T} v_i k_j$$

Para se calcular a relação semântica de pares de fragmentos, compara-se esse vetor resultante de ambos os textos de entrada utilizando a métrica do cosseno supramencionada.

A Figura 2.5 apresenta a arquitetura de funcionamento do algoritmo ESA. Por sua vez, as Figuras 2.6 e 2.7 apresentam o exemplo do processamento de um texto e de um fragmento de texto pelo ESA, respectivamente. Em ambos os casos, as informações de entrada (textos, frases ou termos) são relacionadas com artigos da Wikipédia e posteriormente é construído

um vetor de conceitos com os artigos relacionados a estes. Após comparação vetorial entre ambos os vetores torna-se possível calcular o grau de relacionamento semântico entre as informações de entrada.

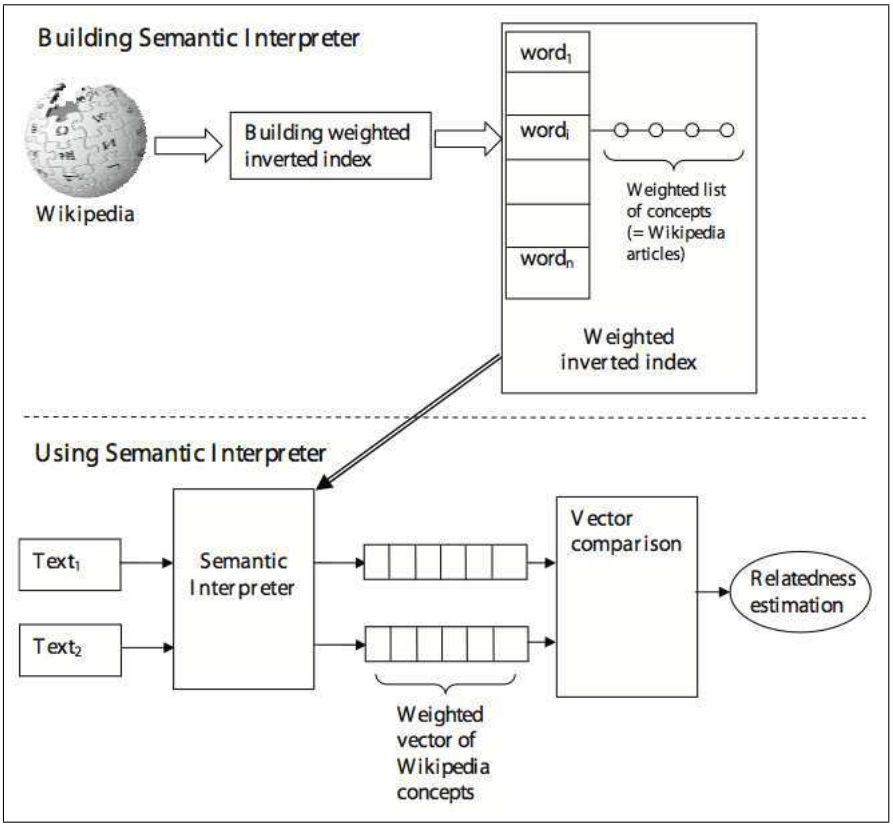


Figura 2.5: Arquitetura de funcionamento do algoritmo ESA [23].

2.8 Projeto RISO-T

O RISO-T[16] representa uma nova abordagem de arcabouço arquitetural, capaz de proporcionar a sistemas de recuperação de informação uma forma eficaz para indexar semanticamente documentos textuais de maneira automática, desambiguando coerentemente suas informações, estabelecendo domínios de contexto, enriquecendo semanticamente os conceitos identificados por meios de fontes externas e heterogêneas de informação, além de proporcionar um ambiente híbrido de consultas, permitindo, assim, tanto consultas sintáticas, quanto semânticas.

[16]RISO-T é o acrônimo de Recuperação de Informação Semântica de Objetos Textuais.

#	Input: *"U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials."*	Input: *"The development of T-cell leukaemia following the otherwise successful treatment of three patients with X-linked severe combined immune deficiency (X-SCID) in gene-therapy trials using haematopoietic stem cells has led to a re-evaluation of this approach. Using a mouse model for gene therapy of X-SCID, we find that the corrective therapeutic gene IL2RG itself can act as a contributor to the genesis of T-cell lymphomas, with one-third of animals being affected. Gene-therapy trials for X-SCID, which have been based on the assumption that IL2RG is minimally oncogenic, may therefore pose some risk to patients."*
1	Iraq disarmament crisis	Leukemia
2	Yellowcake forgery	Severe combined immunodeficiency
3	Senate Report of Pre-war Intelligence on Iraq	Cancer
4	Iraq and weapons of mass destruction	Non-Hodgkin lymphoma
5	Iraq Survey Group	AIDS
6	September Dossier	ICD-10 Chapter II: Neoplasms; Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism
7	Iraq War	Bone marrow transplant
8	Scott Ritter	Immunosuppressive drug
9	Iraq War- Rationale	Acute lymphoblastic leukemia
10	Operation Desert Fox	Multiple sclerosis

Figura 2.6: Exemplo de texto processado pelo algoritmo ESA [23].

#	Input: *"equipment"*	Input: *"investor"*
1	Tool	Investment
2	Digital Equipment Corporation	Angel investor
3	Military technology and equipment	Stock trader
4	Camping	Mutual fund
5	Engineering vehicle	Margin (finance)
6	Weapon	Modern portfolio theory
7	Original equipment manufacturer	Equity investment
8	French Army	Exchange-traded fund
9	Electronic test equipment	Hedge fund
10	Distance Measuring Equipment	Ponzi scheme

Figura 2.7: Exemplo de fragmento processado pelo algoritmo ESA [23].

Em sua estrutura interna, o RISO-T organiza suas informações em mapas de tópicos semânticos, proporcionando várias possibilidades de se obter uma determinada informação.

A seguir, é descrito o projeto RISO-T em 4 módulos, dando maior ênfase ao módulo 3 que é onde se concentra este trabalho de dissertação. A Figura 2.8 apresenta a arquitetura geral do RISO-T e como estão dispostos cada um dos módulos, suas entradas e saídas. Cada módulo será explicado a seguir.

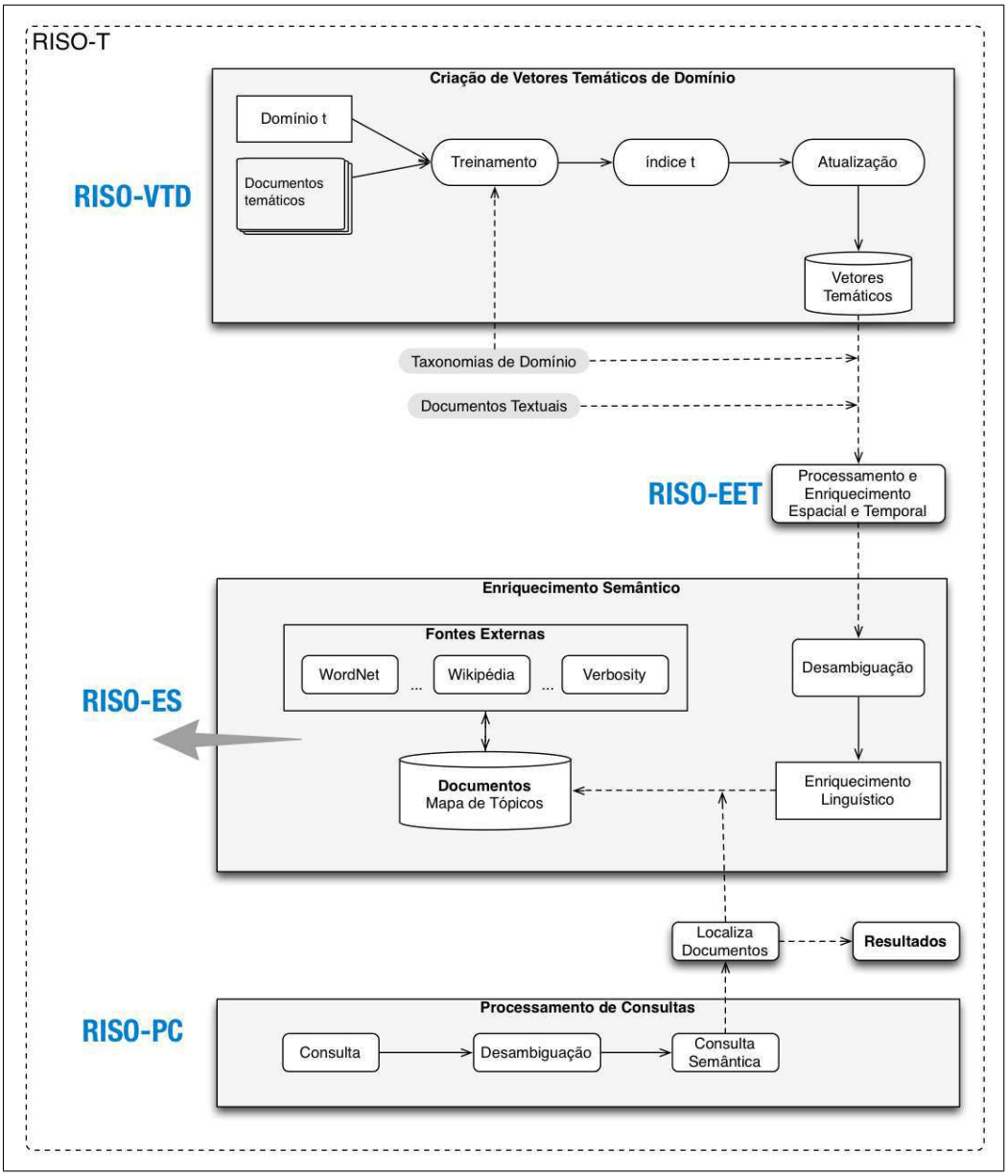


Figura 2.8: Visão Geral do Projeto RISO-T.

2.8.1 RISO-VTD: Criação de Vetores Temáticos de Domínios

Antes do início da indexação de documentos são criados vocabulários típicos dos diversos domínios do conhecimento, denominados Vetores Temáticos de Domínios (VTD). Estes vetores serão utilizados para a classificação dos documentos a serem indexados para auxiliar na desambiguação dos termos contidos neles.

Inicialmente, o sistema obtém da Wikipédia a estrutura taxonômica desejada e artigos correspondentes para realizar o processamento de criação de vetores temáticos de domínios específicos. Por exemplo, um termo como *trigonometria*, provavelmente, far-se-á presente no domínio *Matemática*. Já o termo *célula*, por sua vez, é provável pertencer tanto ao domínio *Biologia*, representado a menor unidade viva dos seres vivos, quanto ao domínio *Matemática* representando o menor unidade de uma tabela. A Figura 2.9 apresenta graficamente o processo de estruturar as informações dos textos dos artigos da Wikipédia em domínios.

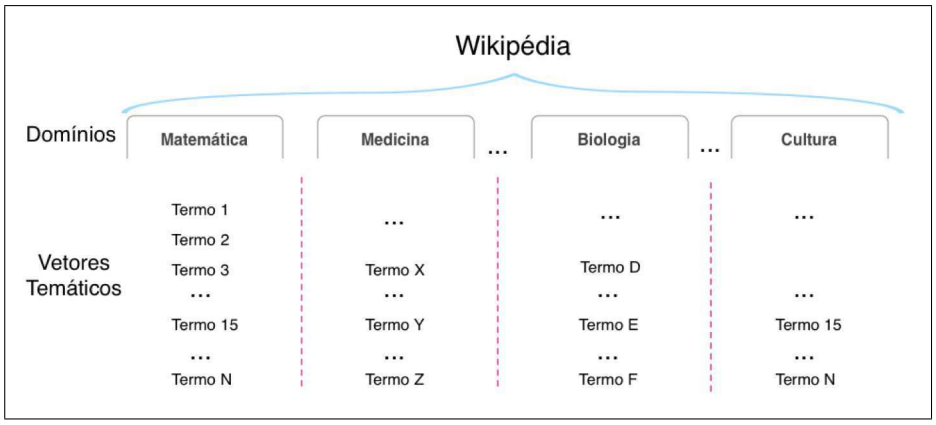


Figura 2.9: Criação de vetores temáticos de domínio baseado nas informações da Wikipédia.

Com a conclusão desta etapa [45], torna-se possível, por meio da comparação vetorial entre os termos de um documento a ser analisado e os termos presentes em cada um dos vetores de domínios, classificá-lo como pertencente a algum domínio específico. Por exemplo, documentos que contenham termos, como: *jurisprudência*, *habeas corpus*, *constitucionalidade*, entre outros, por meio da comparação vetorial serão provavelmente classificados como pertencentes ao domínio *Direito*, uma vez que existirá pouca ou nenhuma correspondência com os vetores temáticos de domínios como *Matemática*, *Biologia*, *Física* e outros.

2.8.2 RISO-ET: Identificação e Tradução Semântica de Conceitos Espaço-Temporais

A etapa de identificação e interpretação semântica de conceitos espaço-temporais visa detectar em documentos textuais, tanto elementos cuja semântica expressem noções de tempo ou espaço, quanto objetos e eventos com uma componente espaço-temporal. Por exemplo, expressões como: *hoje*, *amanhã*, *três meses antes*, *dia da Independência do Brasil*, são exemplos de expressões temporais que devem ser compreendidas pelo sistema e, logo após, substituídas por valores temporais. Já eventos/objetos como "Independência do Brasil" e "SBBD2013", possuem uma validade temporal bem determinada. Por exemplo, *dia da independência do Brasil* equivale a 7 de setembro de 1822, e neste caso, "amanhã" significa a data atual do documento adicionado de mais um dia e "independência do Brasil" é um evento que ocorreu em 7.9.1822. Por sua vez, "SBBD2013" representa um evento do "Simpósio Brasileiro de Bancos de Dados" que será realizado em "Outubro de 2013" na cidade de "Recife - Pernambuco".

O mesmo raciocínio aplicado ao módulo de análise temporal é válido para a etapa espacial. Termos como *Constantinopla*, apesar de não representar nenhuma região física atual do planeta, a partir da análise espacio-temporal proposta pelo RISO-T, relacionará *Constantinopla* com *Istambul* (representante atual), com aspectos temporais e espaciais que se estenderam de 330-1922.

Assim sendo, a análise visa interpretar informações tanto espaciais, quanto temporais presentes em documentos textuais submetidas ao RISO-T, potencializando as possibilidades dos usuários de tal sistema.

2.8.3 RISO-ES: Enriquecimento Semântico de Conceitos, Desambiguação e Indexação Semântica de Documentos

Este módulo tem como finalidade, dentro do RISO-T, promover o enriquecimento semântico de conceitos baseados em informações presentes em fontes externas e heterogêneas ao sistema. Estas informações são desambiguadas para, em seguida, receberem a correta indexação semântica de determinado documento. Os detalhes deste processo serão descritos no capítulo 4.

2.8.4 RISO-PC: Processamento de Consultas e Ambiente de Interatividade

O módulo de consulta visa processar a consulta emitida pelo usuário. Na tentativa de relacionar o texto da busca com conceitos presentes na estrutura semântica proposta. Assim, o usuário terá a possibilidade de desambiguar seus termos enriquecendo-os por meio da adição de novos termos semanticamente relacionados aos originais da consulta. Este módulo, ainda a ser desenvolvido, presente na base da Figura 2.8 apresentará uma estrutura de grafo dinâmico, possibilitando o caminhamento semântico determinado pelo mapa de tópicos da base de documentos.

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta trabalhos relacionados com a abordagem proposta por esta dissertação em vários âmbitos, desde a evolução de sistemas de RI convencionais até a análise de enfoques que envolvem processamento de conceitos e estruturação de conhecimento.

3.1 Retroalimentação de Relevância

Ruthven e Lalmas [55] buscaram aperfeiçoar o processo de recuperação da informação na situação específica em que o usuário conhece os documentos que deseja recuperar e os termos relacionados a eles. Entretanto, o usuário não sabe expressar a consulta para que o sistema consiga processá-la de maneira eficiente.

Diferente de outras abordagens, os autores inovaram com a proposta de uma técnica denominada *retroalimentação da relevância - (Best match)* que, segundo resultados apresentados, melhora de forma significativa a qualidade do retorno dos sistemas de recuperação de informação, a medida que avalia o grau de satisfação do usuário com o resultado obtido e utiliza essa informação para classificar novos resultados, sem se importar com a forma de construção da consulta. A retroalimentação da relevância é uma estratégia que potencializa o processo de recuperação da informação e deve ser vista como um complemento a estes sistemas. Entretanto, esta técnica apresenta algumas fragilidades, dentre elas, a dependência da avaliação coerente do usuário, que influencia diretamente o conjunto de resultados apresentados e a impossibilidade de que elementos que melhorariam o conjunto resposta sejam apresentados, avaliados e redefinidos no *ranking* devido a estarem empatados com zero vo-

tos juntamente com inúmeros outros elementos. Dessa maneira, este enfoque permite, por exemplo, que elementos que melhorariam o conjunto resposta ocupem posições distantes no *ranking* e consequentemente na apresentação ao usuário, o que dificulta sua análise e realocação.

Neste trabalho não foi implementada a técnica de retroalimentação de relevância. Por outro lado, a técnica aqui apresentada, busca melhorar a eficácia do conjunto resposta produzido por sistemas de RI por meio de uma abordagem semântica, retirando o usuário como parte fundamental do processo e de sua qualidade. Todavia, é possível associar em trabalhos futuros as perspectivas da retroalimentação para influenciar de maneira implícita a construção de tesauros e indexação de documentos.

3.2 Expansão de Consultas

Expansão de consulta é uma técnica da RI que proporciona a adição de termos à consulta emitida pelo usuário com o objetivo de ampliar o espaço de busca da consulta e consequentemente sua cobertura.

Ruthven [54] comparou duas técnicas de expansão de consultas: interativa (com participação do usuário) e automática (sem intervenção do usuário). O autor defende que a expansão de consulta interativa constitui uma técnica eficiente, embora esta eficiência esteja diretamente relacionada com a experiência dos usuários que irão informar os termos relevantes para a expansão.

É compreensível que os benefícios desta técnica não serão alcançados se o usuário não possuir a experiência adequada para identificar termos úteis. Este fato torna a expansão automática um caminho interessante quando a experiência do usuário não for um requisito (o que é o mais comum). Notavelmente, problemas semânticos muito recorrentes como a polissemia, ambiguidade e acronímia, podem ser evitados por um usuário experiente, mas não serão evitados por usuários leigos. Palavras como *manga* e *jaguar*, que possuem múltiplas interpretações, por exemplo, podem retornar os mais diversos resultados, muitos deles não relevantes. O usuário leigo não saberá expressar seu objetivo em consulta, e assim sendo, a técnica de expansão automática se torna mais adequada.

O trabalho de Abdelali et al. [1] apresenta uma técnica de expansão semântica de consul-

tas que relaciona a consulta original, emitida pelo usuário, com termos de expansão presentes em uma base de palavras com um contexto determinado. Experimentos apresentados comprovaram uma melhoria significativa na cobertura e na precisão quando comparados com outros mecanismos existentes. Contudo, foram utilizadas coleções de dados em árabe para validação do trabalho o que limita que algumas afirmações sejam aplicáveis ao idioma inglês e às técnicas de expansão de consultas construídas exclusivamente para este idioma.

No trabalho aqui apresentado, foi realizado o enriquecimento semântico de conceitos relacionando automaticamente a expansão de termos da consulta aos elementos da estrutura de mapa de tópicos conceituais previamente construída. Diferente do trabalho de Abdelali, as técnicas de expansão semântica, validação e verificação aqui apresentadas foram projetadas para o idioma inglês, haja vista ser o idioma mais difundido no mundo. Além disso, diferente de Ruthven e Abdelali, a expansão semântica proposta por este trabalho ocorre em dois momentos: durante a indexação do documento - onde para cada conceito identificado no documento é realizada a expansão semântica para posterior indexação do documento - e no momento de relacionar a consulta do usuário com conceitos previamente estruturados, ampliado a possibilidade de relacionar os termos da consulta com documentos semanticamente relacionados. Por fim, outra diferença entre a expansão de termos da consulta proposta por este trabalho e os demais trabalhos citados é a obtenção dinâmica de conceitos envolvendo várias fontes distintas e heterogêneas - dicionários, enciclopédias e de sentido comum. Nos demais casos analisados, as informações utilizadas para a expansão são limitadas a apenas uma instância de uma dessas fontes.

3.3 Sistemas de RI Auxiliados por Formulários

Allan [4] propôs uma técnica para melhorar a precisão dos sistemas de recuperação de documentos utilizando informações obtidas do usuário por meio do preenchimento dos chamados *formulários de melhor compreensão*, que eram gerados como informação adicional sempre que determinada consulta não fosse plenamente processada pelo sistema. O engenho de busca utiliza estas respostas - informação adicional - para melhorar o processo de recuperação da informação. Todavia, o preenchimento destes formulários por parte do usuário é uma tarefa onerosa que consome tempo e energia e muitos não estão dispostos a realizar, o que

dificulta a utilização efetiva desta técnica.

A criação da estrutura semântica, sugerida nesta dissertação, substitui a necessidade dos formulários de melhor compreensão, sendo que, a informação adicional necessária é obtida por meio dos relacionamentos semânticos entre conceitos presentes na estrutura dos mapas de tópicos previamente construídos.

3.4 Relações Semânticas Utilizando Wikipédia

Milne [46] propôs uma nova técnica para encontrar relações semânticas existentes entre termos e, dessa maneira, desambiguá-los de forma mais eficiente, proporcionando, por consequência, melhor qualidade da informação recuperada. Diferente de outros trabalhos que buscam inferir relações semânticas entre termos incluindo o conteúdo presente nos artigos da Wikipédia, a inovação de Milne consiste em utilizar unicamente a estrutura de *links* e títulos dos artigos da Wikipédia sem a necessidade de processar o conteúdo textual de cada artigo.

Por utilizar apenas a informação de *links* e títulos dos artigos presentes na Wikipédia para inferir relações semânticas, o autor conseguiu um processo capaz de melhorar a eficiência tanto no desempenho quanto no armazenamento de informações. Contudo, é possível melhorar a precisão proporcionada por esta arquitetura de solução incluindo-lhe outros critérios como, por exemplo, análise de palavras-chave.

Assim como o trabalho proposto por Milne, a abordagem aqui apresentada também realiza desambiguação e utiliza as informações da Wikipédia para o enriquecimento de conceitos. Além disso, outras fontes de informação tais como WordNet, Verbosity, Wiktionary e Reverb também são consultadas e interoperadas ampliando a capacidade de identificação e processamento de termos propostos no trabalho de Milne.

3.5 Validação de Algoritmos de RI

Callan et al. [7] propuseram um método para construir e estabelecer corpus genéricos aptos a realizarem validação de algoritmos e de metodologias propostas para a área de recuperação de informação. Inovaram provendo a criação de uma coleção de dados ampla contendo diversos temas, gêneros e formatos com o objetivo de forçar os algoritmos na área de RI a

serem mais genéricos e que a validação/verificação destes algoritmos fosse padronizada e não ad-hoc.

Nesse trabalho, Callan et al. criticam a validação de algoritmos e metodologias que utilizaram coleções de documentos com temas e gêneros restritos e obtiveram bons resultados em suas hipóteses.

Infelizmente, a validação de trabalhos que envolvem relações semânticas não constitui uma tarefa trivial. Além de ser influenciada por fatores culturais, a necessidade de análise de especialistas ou a produção de um corpus coerente e sem viés são atividades onerosas; além disso, abordagens do tipo *face validity* [1], nem sempre representam toda a heterogeneidade necessária e podem ser realizadas com os recursos disponíveis.

He e Ounis [34] apresentaram uma técnica para avaliar a qualidade da expansão de consultas que serão utilizadas na recuperação de informação. A inovação provém da utilização de dois fatores para medir a qualidade da expansão de consultas: um deles é o contexto geral que se refere em linhas gerais ao assunto tratado pelos documentos de interesse do usuário e o outro, a qualidade da consulta fornecida pelo usuário.

Neste trabalho parte da validação foi norteada pela proposta de Callan et al., dessa forma incluiu-se fontes de conceitos que tratavam de variados temas, gêneros e formatos buscando não enviesar a medição para conjuntos de conceitos em que a estrutura é mais robusta.

3.6 Relacionamento de Termos da Consulta com Conceitos da Wikipédia

Egozi et al. [18] criaram um novo algoritmo que recupera documentos relevantes analisando, além das palavras-chave presentes nas consultas, suas possibilidades de interpretação. Muitos dos sistemas de RI tentam selecionar documentos que possuem exatamente os termos presentes na consulta, e isso, por vezes, resulta ineficiente, visto que, para consultas difíceis[2] se faz necessário uma compreensão semântica da consulta. Tal conhecimento pode ser ob-

[1] Estratégia de validação que consiste em utilizar a opinião de observadores para comprovar ou refutar alguma hipótese.

[2] Consultas difíceis são aquelas cujo objetivo não estão totalmente explícitos apenas realizando uma análise léxica; se não, aquela que se faz necessária uma interpretação semântica da informação.

tido por meio de possíveis relações de sentido entre os termos da consulta e seus diversos significados.

Comumente, essas informações são desprezadas, porque os algoritmos convencionais estão preparados apenas para buscar por palavras exatas. Foi proposta pelos autores, a construção de um algoritmo de aprendizagem supervisionada capaz de combinar conceitos presentes na Wikipédia a termos da consulta emitida pelo usuário e dessa maneira melhorar tanto a cobertura quanto a qualidade do conjunto resposta produzido por sistemas de RI, rompendo a restrição de apenas buscar por palavra-chave. Contudo, esta abordagem não descreve se é realizada desambiguação entre os termos presentes na consulta emitida pelo usuário e os conceitos da Wikipédia, podendo assim, ocorrer uma associação equivocada entre termos e conceitos o que acarretará em um conjunto resposta inútil ao usuário. Além disso, o fato de utilizar apenas as informações conceituais da Wikipédia faz com que termos presentes na consulta que não existam nesta fonte de informação não sejam úteis para promover os benefícios propostos pela técnica, sendo este um fator limitante.

Os experimentos apresentados utilizaram os dados do TREC e confirmaram a superioridade do método em relação ao estado da arte.

A abordagem sugerida nesta dissertação implementa técnicas de desambiguação, associando de maneira coerente termo e conceito. Além disso, são utilizadas várias fontes de informação em conjunto com a Wikipédia, dentre elas, Wiktionary, Reverb, Verbosity e WordNet no intuito de ampliar as possibilidades de identificar os elementos da consulta e associá-los a conceitos de modo a melhorar a qualidade do conjunto resposta produzido pelo sistema de RI que implementar este enfoque.

3.7 Junção de Fontes de Informação

Suchanek et al. [60] apresentam um sistema de software denominado YAGO[3], visando a construção automática de ontologias por meio da junção de fontes de conhecimento, como forma de resolver parte dos problemas apresentados anteriormente. Diferente da abordagem apresentada nesta dissertação, o trabalho de Suchanek et al., não utiliza a ontologia construída para realizar desambiguação de termos e indexação de documentos, seu único

[3]http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

objetivo é fornecer e enriquecer tal estrutura. Além disso, a quantidade e heterogeneidade das fontes de informação de conteúdo semântico no trabalho aqui apresentado tais como: *Wikipédia*[4]*, WordNet*[5] *e Wiktionary*[6] é superior ao trabalho de Suchanek.

3.8 Indexação Semântica

O trabalho proposto por Calo Abi et al. [10] apresenta um sistema de suporte a indexação de documentos que sugere, de maneira didática, um conjunto de tópicos e palavras-chave relevantes para serem utilizadas no momento de se classificar determinado documento. Tal método de indexação, usa as categorias da Wikipédia como taxonomia conceitual, construindo com ela um grafo acíclico dirigido. De posse do documento a ser indexado, o algoritmo consegue sugerir termos de interesse ao bibliotecário, e posteriormente relaciona estes termos com categorias da Wikipédia.

Propriedades de grafos são utilizadas para adicionar pesos (ponderar) conceitos. O objetivo principal desta investigação foi o de encontrar rapidamente os principais tópicos de determinado documento para serem posteriormente utilizados métodos de indexação estatística, tais como: TF-IDF[7] e LSA[8] [17]. Em oposição ao que é proposto por este trabalho de dissertação, o trabalho proposto por Calo Abi não relaciona termos extraídos dos documentos entre si, mantendo apenas a informação taxonômica das categorias da Wikipédia como relacionamento semântico entre conceitos e documentos.

Para este trabalho de dissertação, cada documento fornecido como entrada no sistema é inicialmente processado para a identificação de conceitos que posteriormente são enriquecidos semanticamente e estruturados em mapas de tópicos semânticos. Durante esta etapa, inúmeras relações semânticas são criadas entre os termos presentes nos documentos e outros termos de outros documentos já presentes na estrutura. A indexação do documento ocorrerá

[4]http://www.wikipedia.org/

[5]http://wordnet.princeton.edu/

[6]http://www.wiktionary.org/

[7]Frequência do termo – frequência inversa de documento (ou seja, a frequência de ocorrência do termo na coleção de documentos). O TF-IDF determina o peso do termo em um documento.

[8]Análise Semântica Latente é uma técnica de processamento de linguagem natural que analisa a relação entre o conjunto de termos e o conjunto de documentos por meio de uma técnica matemática de decomposição de valor singular (DVS).

apenas posteriormente, após serem realizadas todas as desambiguações necessárias de modo a associar de maneira correta o documento aos conceitos nele contidos.

3.9 Análise Formal de Conceitos

O trabalho proposto por Kumar [40] apresenta uma abordagem para modelagem de conceitos por meio da Análise Formal de Conceitos. Esse tipo de modelagem conceitual apresenta grande potencial de organização do conhecimento em uma hierarquia explícita, apresentando um arcabouço matemático capaz de oferecer uma representação conceitual do conhecimento por meio de uma estrutura semelhante a uma grande tabela, em que linhas são conceitos e colunas característica de conceitos.

A vantagem dessa abordagem é o formalismo matemático que alicerça todas as operações de busca e recuperação de informação. Entretanto, para uma grande quantidade de conceitos ou características, existe a necessidade de se realizar operações onerosas sobre a tabela o que sobrecarrega recursos de armazenamento e torna as operações de consulta e recuperação lentas. Por isso apenas parte deste formalismo foi considerado.

3.10 Projeto UBY

Gurevych et al. [31] propuseram o UBY[9] que trata-se de um recurso para o processamento de linguagem natural, baseado na norma ISO LMF[10], que combina uma vasta gama de informações construídas tanto de maneira colaborativa quanto por especialistas. O projeto UBY disponibiliza estrutural e semanticamente versões interoperáveis de 10 fontes de informação em duas línguas (Inglês e Alemão): WordNet, Wikitionary, Wikipedia-EN[11], FrameNet, VerbNet, OmegaWiki, IMSLex-Subcat, GermaNet, Wikitionary e Wikipedia-GE[12], além disso, ele foi projetado para ser diretamente extensível por novas linguagens e recursos por meio de sua API.

[9] http://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/

[10] http://lexicalmarkupframework.org/

[11] Informações e artigos da Wikipédia para o idioma inglês.

[12] Informações e artigos da Wikipédia para o idioma alemão.

3.11 Desafios da RI

Belkin [5] expôs os grandes desafios da RI moderna. Um dos principais problemas abordados refere-se à criação de um sistema ideal que seja capaz de proporcionar uma melhor interação com o usuário, melhorando consequentemente sua usabilidade, utilidade e satisfação.

Segundo o autor, o usuário deve ser considerado como elemento fundamental dentro do processo de recuperação de informação e não apenas como um mero agente passivo dentro deste contexto. Este trabalho contradiz muitas das ideias de Ruthven [54], comentado anteriormente. A grande debilidade em se utilizar o usuário como parte do sistema de recuperação de informação é que, dependendo do grau de experiência deste usuário, as informações fornecidas por ele ao sistema podem intervir negativamente para uma recuperação eficiente, ou seja, alguns usuários não serão capazes de fornecer informação útil ao sistema.

3.12 Comentários Finais

Não foi encontrado nenhum trabalho que explore as diversas relações semânticas existentes na linguística para enriquecer tanto os índices conceituais de um documento, quanto os termos de uma consulta. Por outro lado, com a inexistência de trabalhos com os mesmos propósitos, a atividade de validação *offline* se torna difícil, sendo necessário recorrer a trabalhos semelhantes como o UBY, por exemplo, embora não tenham sido projetados com o mesmo objetivo.

Capítulo 4

RISO-ES: Enriquecimento Semântico, Desambiguação e Indexação

Atualmente, grande parte dos engenhos de busca textual estudam estratégias para organizar e ranquear de maneira mais eficiente o conteúdo recuperado. Processos como a retroalimentação de relevância [48; 55] implementada pelo google+, por exemplo, apresentam uma vertente predominante, mas, inalterando a maneira de como é realizado o processo de indexação e recuperação de informação. Para uma grande base de informação, com centenas de milhares de elementos indexados, a possibilidade de um elemento da consulta não ser encontrado, tende a ser muito pequena; muito embora a qualidade dos resultados fornecidos seja questionável.

Quando um usuário emite uma consulta como *"luva"*, por exemplo, pode estar querendo comprar, vender, ou se informar sobre algum tipo de luva, que pode ser para o *frio*, *cirúrgica*, *dinheiro (pago aos atletas)* ou para prática de esportes que, por sua vez, pode ser *baseball*, *futebol*, *boxer*, *rapel*, entre outros.

As abordagens convencionais recuperam todas as informações indexadas por *"luva"* independentemente do tipo de interesse do usuário. Diferente da abordagem convencional, a proposta deste trabalho é a de determinar informações conceituais e estabelecer relações semânticas entre esses conceitos extraídos de documentos, para proporcionar ao usuário um melhor filtro e flexibilidade em suas pesquisas. A Figura 4.1 apresenta graficamente o exemplo comentado, estruturado por este trabalho de dissertação.

O valor de uma biblioteca digital não reside apenas nos documentos em si ou na ca-

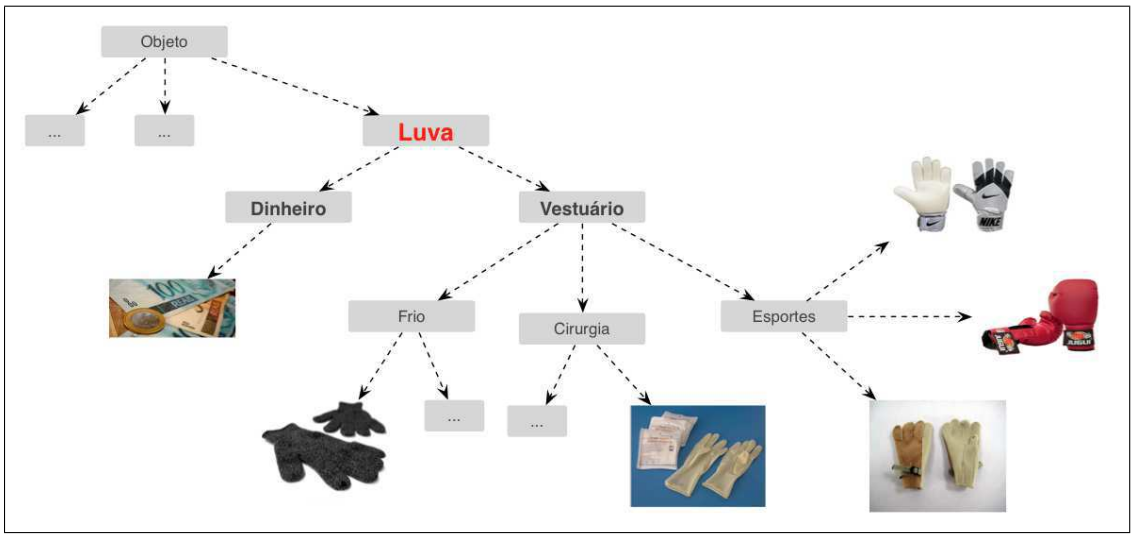


Figura 4.1: Exemplo da estruturação da instância *"luva"* no projeto RISO-ES.

pacidade de consultá-los e acessar o acervo livremente pela Web, mas sim, na riqueza de informações providas por estes documentos, sua ampla gama de entidades, eventos e tópicos relacionados. Acredita-se que a extração e organização de informações de maneira conceitual possa permitir a recuperação e apresentação de informações mais relevantes e, com isso, melhorar, não apenas a maneira como engenhos de busca identificam e ranqueiam conteúdo, como também, permitir o enriquecimento de critérios de busca orientados por atributos.

Por óbvio, a atividade de se construir e manter sistemas semânticos apresenta inúmeros desafios, ao passo que sugere a possibilidade de criar poderosas aplicações e paradigmas de descoberta de informação. Explorar conceitos em documentos textuais para criar automaticamente uma visão semântica agregada de toda informação disponível é potencialmente útil não apenas para recuperação de informação, como também, para outras áreas, tais como, a Web Semântica. Nesta área, a construção de ontologias geralmente exige inúmeros recursos, tais como, tempo e esforço de pessoas capacitadas, tornando comumente oneroso o desenvolvimento de projetos com este fim que, na maioria das vezes, desencoraja sua construção.

Para o enfoque aqui apresentado, diferente de outras abordagens, um conceito constitui o átomo de significado, a menor unidade de informação capaz de expressar de maneira completa determinado sentido ou ideia, seja ela real ou abstrata. Neste enfoque, o conceito é constituído pela fusão de três tipos de informações: *informações de dicionário* (termos que refletem seu papel primeiro diante da língua em consideração), *informações enciclopédicas*

(informações de conhecimento de mundo) e, por fim, *informações de sentido comum* (o veredito de um grupo de pessoas sobre instâncias e sua função diária e pragmática). A Figura 4.2 apresenta um exemplo da criação e representação conceitual proposta por este trabalho de dissertação, no qual, a composição conceitual envolve informações oriundas de fontes de dicionários (WordNet), enciclopédias (Wikipédia) e sentido comum (Verbosity), ao lado, encontra-se uma instância exemplificadora de tal estrutura. A Figura 4.3 apresenta uma visão geral do processo de indexação semântica de documentos textuais.

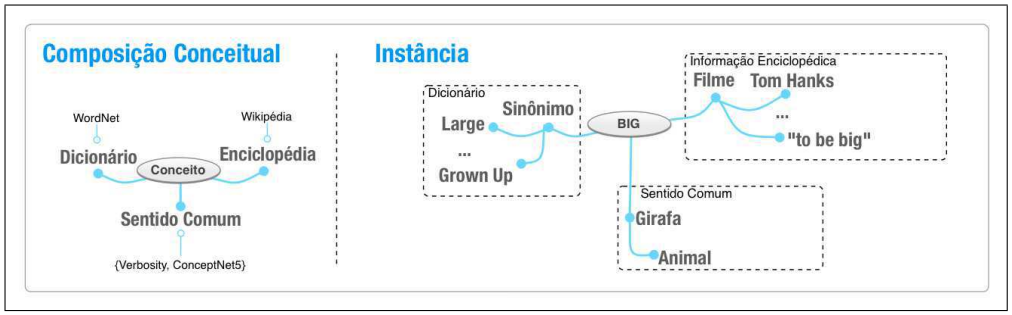


Figura 4.2: Estrutura conceitual do sistema RISO-ES e instância exemplificadora.

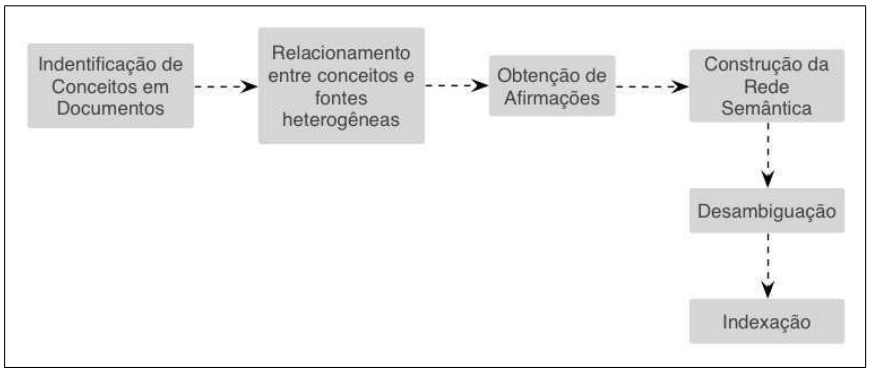


Figura 4.3: Fluxo das etapas para indexação semântica de conceitos.

4.1 Metodologia de Trabalho

A metodologia de trabalho proposta consistiu de seis fases. A seguir, apresenta-se uma síntese de cada fase, uma vez que o processo de solução detalhado será descrito no capítulo 4. Inicialmente, deu-se maior foco à análise de trabalhos relacionados à construção automática

de mapas de tópicos e outras estruturas que vinculassem fontes de informação, tais como, *Wikipédia* e *WordNet* visando o enriquecimento de conceitos que serão descritos em detalhes no capítulo 2.

4.1.1 *Primeira fase:* Obtenção de Conceitos em Documentos Textuais

Sejam:

- D : a coleção contendo *todos os documentos textuais* de uma *biblioteca digital*[1].
- $c(x)$: a função que recebe um documento x como parâmetro e retorna o conjunto de conceitos presentes no documento x.

Então, a *primeira fase* deste trabalho consistirá em obter:

$$C_D = \{c(d), d \in D\} \tag{4.1}$$

O processo contemplado por esta seção, busca identificar e extrair conceitos presentes em todos os documentos da coleção, previamente marcados, utilizando o algoritmo de detecção de conceitos em documentos proposto por [45] que, basicamente, consiste de uma variante para a técnica de marcação de sintagmas proposta por [51]. Ou seja, baseado em vetores temáticos de domínios, identificam-se sintagmas nominas e verbais que são combinações de unidades linguísticas que podem formar orações (nominais ou verbais) e que apresentam características de prováveis conceitos, sendo úteis para anotá-los posteriormente.

Todos os sintagmas marcados pelo algoritmo, são considerados conceitos de forma impreterível.

4.1.2 *Segunda fase:* União de Fontes Heterogêneas de Informação

Nesta fase, busca-se unir em um único ambiente todas as fontes de informação externas ao RISO-ES: *WordNet, Wikipédia, Wiktionary, DBPedia, Conceptnet5, Verbosity e Reverb*. Informações de dicionários, enciclopédias e de sentido comum são estruturadas de maneira centralizada e uniforme, favorecendo assim, a utilização destas informações no processo de identificação de conceitos presentes em documentos e seu adequado enriquecimento semântico.

[1] A seção 2.6 apresenta uma definição de bibliotecas digitais.

4.1.3 *Terceira fase:* Enriquecimento Semântico de Conceitos

Para Conceitos Perfeitamente Identificados

Conceitos identificados que possuem correspondência exata, ou seja, os mesmos caracteres dispostos na mesma ordem, em alguma das fontes de informação utilizadas pelo RISO-ES, são enriquecidos diretamente sem a necessidade de nenhum processamento complementar.

Por exemplo, para o conceito *"car"*, tem-se:

- *Wiktionary* $\rightarrow$ *Sinonímia* $\rightarrow$ *auto*.
- *WordNet* $\rightarrow$ *Hiponímia* $\rightarrow$ *motor vehicle*.
- *WordNet* $\rightarrow$ *Meronímia* $\rightarrow$ *door, air bag, roof*.
- *Verbosity* $\rightarrow$ *Hiperônimo* $\rightarrow$ *personal vehicle*.

Para Conceitos Parcialmente Identificados

Conceitos parcialmente identificados são conceitos onde apenas parte das palavras que o constitui estão presentes em algumas das fontes de informação adotadas. Para estes conceitos deve-se realizar um processamento complementar calculando inicialmente sua proximidade com conceitos presentes nas fontes de informação e uma vez selecionado o conceito mais próximo (maior quantidade de palavras consecutivas presentes no conceito), estabelece-se com este uma relação de hiponímia/hiperonímia.

Por exemplo, caso seja obtido o conceito *"camaro 2010"* em algum documento, mas em nenhuma das fontes de informação exista exatamente esse conceito, haverá a necessidade de serem realizados novos processamentos na tentativa de identificar nas fontes externas, conceitos "próximos", nesse sentido:

$$\{camaro, 2010, camaro\ 2010\ mpg, \ldots\}$$

seriam prováveis candidatos por divergirem do conceito originalmente buscado por apenas um único sintagma, para mais ou para menos como no caso de *"camaro 2010 mpg"* e *"camaro"* respectivamente.

Em casos onde sejam encontrados nas fontes de informação conceitos com alguma correspondência sintática com partes do conceito que se deseja enriquecer, estabelece-se uma relação de hiper/hiponímia uma vez que as únicas possibilidades para este caso são: o conceito selecionado nas fontes de informação é mais específico do que o conceito que se deseja enriquecer, ou o contrário. Assim sendo, para o exemplo apresentado, seriam estabelecidas as seguintes relações:

- *camaro 2010* $\rightarrow$ *É um* $\rightarrow$ **camaro**
- *camaro 2010* $\rightarrow$ *feito Em* $\rightarrow$ **2010**
- **camaro 2010 mpg** $\rightarrow$ *É um* $\rightarrow$ camaro 2010

A partir da relações estabelecidas com este conceito presente nas fontes de informação, torna-se possível dar seguimento ao processo de enriquecimento semântico.

4.1.4 *Quarta Fase:* Construção Automática de Ontologias

Para a ciência da computação uma ontologia[2] é um modelo de dados que representa um conjunto de conceitos de um domínio e relacionamentos entre estes. Durante o processo de obtenção de informações conceituais das fontes externas ao RISO-T, ocorre uma estruturação conceitual, que consiste em analisar todas as afirmações (*asserts*) obtidas sobre determinado conceito, tratando e anotando-as semanticamente em formato de ontologias para posteriormente serem desambiguadas e utilizadas na indexação dos documentos. Para esta estruturação, utilizou-se um *framework* de software específico denominado Jena[3], que permite a manipulação de ontologias a partir da linguagem Java, favorecendo a realização de diversas operações de consultas.

A vantagem da estruturação de informações em formato de ontologias é a possibilidade de interoperar com outras ontologias, sendo possível fortalecer bases de ontologias com o conhecimento extraído das fontes externas de informação, como também, ser melhorado por ontologias detalhadas e específicas de algum tema em que o enriquecimento semântico seja pobre. A Figura 4.4, apresenta um exemplo da estruturação de informações em ontologias dos conceitos “sport team” e “san diego museum of man”.

[2] A seção 2.5 apresenta uma discussão mais ampla sobre ontologias.

[3] http://jena.apache.org/

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:coneitos="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rel="http://lsi.dsc.ufcg.edu.br/riso.owl#/r/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/sport_team">
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
 </rdf:Description>
 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/san_diego_museum_of_man">
   <rel:AtLocation rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/balboa_park/n/San_Diego"/>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
 </rdf:Description>
```

Figura 4.4: Estruturação de informações em ontologias.

4.1.5 *Quinta Fase:* Desambiguação e Indexação de Conceitos

Para determinados sintagmas, etiquetados pelo algoritmo proposto por [45], é provável de que estes sejam passivos a mais de uma interpretação. Para estes casos, se faz necessária uma etapa de desambiguação, para que os documentos que o contém sejam indexados ao conceito correto.

Por exemplo, para um determinado documento que possui o sintagma *"jaguar"* etiquetado, após enriquecimento semântico, existirão relacionamentos para este elemento tanto no sentido do *"jaguar animal"* relativo a {*mamífero*, *felino*, *veloz*, *predador*, ...}, quanto no sentido de *"jaguar carro"* relacionado com { *carro luxuoso*, *esportivo*, *raro*, *caro*, *importado*, ... }, entre outras associações.

A etapa de desambiguação busca analisar as possíveis interpretações de sintagmas identificados, juntamente com os trechos do documento onde são encontrados, para então, determinar conceitualmente, onde o documento deve ser indexado.

4.1.6 *Sexta fase:* Avaliação e Comparação da Ferramenta

A Figura 4.5 apresenta uma síntese da processo de validação proposto por este trabalho.

Foi proposto que a validação desse trabalho ocorresse da seguinte maneira:

1. Buscar ferramentas de propósito semelhante ao módulo de enriquecimento semântico do projeto RISO-T (RISO-ES) para realizar a comparação de requisitos tais como: cobertura (quantidade de conceitos identificados), conectividade (para cada conceito identificado, a quantidade de conexões que este estabelece com outros conceitos), qualidade do enriquecimento semântico (se as conexões estabelecidas entre os conceitos

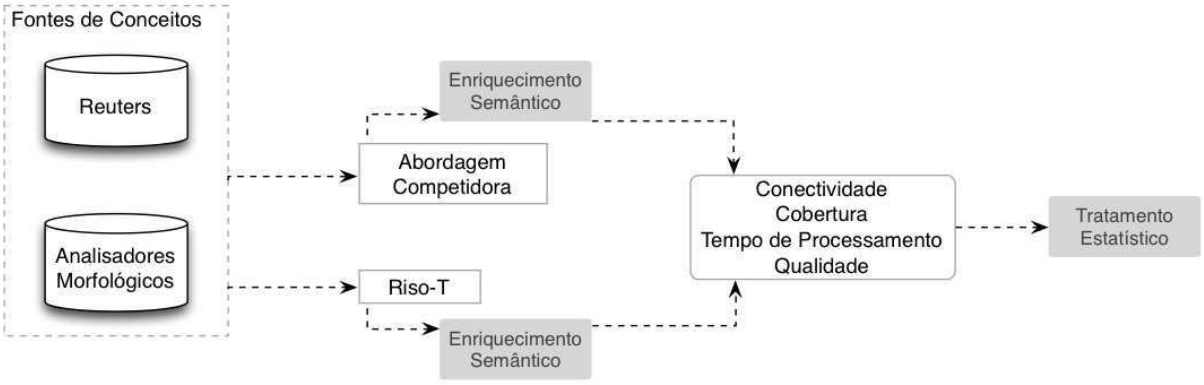


Figura 4.5: Modelo de avaliação proposto por este trabalho.

condizem com a realidade) e tempo de processamento utilizado para calcular cada um dos requisitos.

2. Utilizar as sub-coleções de documentos Reuters-21578 de nomes próprios, regiões geográficas, siglas e acrônimos para serem submetidas às ferramentas do item 1.

3. Utilizar analisadores morfológicos, para incrementar a quantidade de conceitos mencionados em 2, e submeter as ferramentas escolhidas no item 1 para calcular as métricas propostas ainda no item 1.

4. Favorecer o enriquecimento semântico das coleções propostas pelos itens 2 e 3 nas abordagens selecionadas pelo item 1.

5. Calcular as métricas propostas em 1 para cada abordagem selecionada.

6. Realizar testes estatísticos para validar as estratégias propostas para o RISO-ES.

7. Apresentar textual e graficamente os resultados obtidos e analisados.

4.2 Formalização do Problema

A seguir, apresenta-se a formalização do problema da determinação de relações semânticas entre conceitos.

Sejam:

T: o conjunto de sintagmas de um idioma I.

C: o conjunto de todos os conceitos.

$R_c : T \rightarrow \mathcal{P}(C)$: a função que associa a cada sintagma $t \in T$ o conjunto de conceitos representados por t, onde $\mathcal{P}(\mathcal{C})$ é o conjunto das partes de C.

$R_t : C \rightarrow \mathcal{P}(T)$: a função que associa a cada conceito $c \in C$ o conjunto de sintagmas utilizados para representar c.

Por exemplo:

- $R_c(manga)$ = {"parte do vestuário que cobre o braço","fruto da mangueira", "goleiro titular do Brasil na copa de 1966"}
- R_t(*Manihot utilissima*) = {"mandioca", "aipim", "macaxeira"}

Quando $\mid R_c(t) \mid > 1$, t é um homônimo e, nesse caso, espera-se que a ambiguidade da ocorrência de t em um texto específico possa ser removida por meio de um conjunto de sintagmas semanticamente relacionados a t.

D_t: um conjunto de sintagmas utilizados para desambiguar parcialmente um determinado t.

D_t^*: um conjunto de sintagmas utilizados para desambiguar completamente um determinado t, ou seja, D_t^* apresenta um contexto para t que é o assunto ou tema a que t se refere.

Supõe-se, então, que todo sintagma t pode ser desambiguado por um conjunto D_t^* sendo que o par (t, D_t^*) representa um conceito único.

$R_d(t, D_t) = \{c_1, c_2, \ldots, c_z\}$: função que associa sintagmas contextualizados (t, D_t) a conceitos. Sendo que, $R_d(t, D_t^*) = c$ é uma função que relaciona o sintagma t desambiguado ao conceito correspondente.

- $R_d(manga, \{camisa\}^*)$ = {"parte do vestuário que cobre o braço"}.
- $R_d(manga, \{fruta, alimento\}^*)$ = {"fruto da mangueira"}.
- R_d(*S.Paulo,{geográfico, cidade}**) = {Cidade de S.Paulo}.
- R_d(*S.Paulo,{geográfico}*) = {Cidade de S.Paulo, Estado de São Paulo}.

  Apesar de haver diminuído a ambiguidade e evitar que "S. Paulo" fosse interpretado como "time de futebol", "santo", entre outros, o conjunto de sintagmas desambiguadores não foi suficiente, neste caso, para desambiguá-lo completamente.

Existem diversas relações semânticas entre conceitos que podem ser exploradas para enriquecer uma consulta e atender melhor às necessidades de informações do usuário de um sistema de RI.

Como notação genérica para uma relação semântica, define-se:

S: o conjunto de todas as relações semânticas.

Para cada relação semântica $s \in S$, tem-se uma relação $R_s \subseteq C \times C$ de todos os pares de C que estão relacionados por s. Com isso, é possível definir uma função que para cada s em S e c em C determina o conjunto dos conceitos relacionados a c por s:

$$R_{sem} : S \times C \rightarrow \mathcal{P}(C) | R_{sem}(s, c) = \{c_i \in C | < c, c_i > \in R_s\}$$

Sendo o conceito c do par $< s, c >$ determinado por um processo de desambiguação de um termo t, como $R_d(t, D_t^*) = c$, a função R_{sem} determina os conceitos semanticamente relacionadas ao conceito c adequado a um termo t que ocorre em um documento d. Desta maneira estabelece-se:

$$R_{sem}(s, R_d(t, D_t^*)) = \{c_1, c_2, \ldots, c_m\}$$

O conjunto de conceitos relacionados a c representado por um termo t em um documento d por meio de s. Com esta função fecha-se o ciclo de, dado um termo t qualquer, contido em um documento d, determinar-se, para cada relação semântica s, todos os conceitos semanticamente relacionados ao conceito que t representa no documento. Estes novos conceitos serão úteis para enriquecer a indexação de documentos e o processamento de consultas, provendo melhor precisão e abrangência nos documentos recuperados.

Como exemplos, tem-se:

- R_{sem}(*hipônimo*, R_d(*banco de dados*, {*computação*})) = {"banco de dados temporal", "banco de dados espacial"}.
- R_{sem}(*acrônimo*, $R_d(OCL$, {*política, França*})) = {"Organization Communiste Libertaire"}.
- R_{sem}(*merônimo*, $R_d(manga$, {*vestuário*})) = {"camisa"}.
- R_{sem}(*sinônimo*, $R_d(aipim$, {*alimentação, tubérculo*})) = {"macaxeira", "pão de pobre"}.

Cada uma das relações semânticas terá diferentes propriedades relacionais, como reflexividade, assimetria, transitividade, entre outra.

O propósito deste trabalho é transformar sintagmas em conceitos, tanto na indexação de documentos, quanto no processamento de consultas. Dessa maneira, problemas de polissemia estarão resolvidos. Após a transformação, o conceito obtido será enriquecido por meio de relações semânticas.

Para cada documento F considera-se:

- *ind-sin*$(F) = \{t_1, t_2, \ldots, t_n\}$ o conjunto de sintagmas contidos em F.
- *context*$(t_i) = D^*_{t_i}$, a função que determina o conjunto de sintagmas capazes de desambiguar completamente (contexto) o sintagma t_i dentro do parágrafo em que este ocorre em F.

Transformam-se agora os termos sintáticos extraídos em conceitos:

ind-conceito$(F) = \{R_d(t_1 : context(t_1)), R_d(t_2 : context(t_2)), \ldots, R_d(t_n : context(t_n))\} = \{R_d(t_1 : D^*_{t_1}), R_d(t_2 : D^*_{t_2}), \ldots, R_d(t_n : D^*_{t_n})\} = \{c_1, c_2, \ldots, c_n\}$

A mesma função *ind-conceito(F)* pode ser aplicada aos termos de uma consulta. Neste caso, o contexto deve ser obtido do usuário.

Para uma dada consulta Q considera-se:

- $\{q_1, q_2, \ldots, q_m\}$: o conjunto de sintagmas contidos em Q.
- D^*_Q: o contexto informado pelo usuário.
- *ind-conceito*$(q_1, q_2, \ldots, q_m) = \{q_1 : D^*_Q, q_2 : D^*_Q, \ldots, q_m : D^*_Q\} = \{R_d(q_1, D^*_Q), R_d(q_2, D^*_Q), \ldots, R_d(q_m, D^*_Q)\} = \{c_1, c_2, \ldots, c_m\}$

Uma vez determinados os conceitos associados a um documento ou a uma consulta, eles podem ser enriquecidos.

Seja $\theta(c)$: a função que, dado um conceito c, aplica a relação R_{sem} para cada elemento de S, tal que:

$$\theta(c) = \{R_{sem}(s_1 : c) \cup R_{sem}(s_2 : c) \cup \ldots \cup R_{sem}(s_n : c)\}$$

Formalizando o processamento de uma consulta convencional em contraste com uma consulta no ambiente RISO, tem-se:

Seja $rec(t)$ o conjunto de objetos recuperados relacionados a t.

Seja $rec^*(t)$ o subconjunto de $rec(t)$ que contêm apenas objetos úteis para o usuário.

Para uma consulta $Q = \{q_1, q_2, \ldots, q_n\}$ tem-se:

$$rec_{Convencional}(Q) = rec(q_1) \cup rec(q_2) \cup \ldots \cup rec(q_n)$$

No ambiente RISO, o processamento da consulta será:

$$rec_{RISO}(Q) = rec_{RISO}(\{q_1, q_2, \ldots, q_n\})$$

$$rec_{RISO}(\{q_1, q_2, \ldots, q_n\}) = rec_{RISO}(\{\theta(R_d(q_1, D_Q^*)), \theta(R_d(q_2, D_Q^*)), \ldots, \theta(R_d(q_m, D_Q^*))\})$$

Ou seja, um termo q de uma consulta Q, antes de ser processado pela função (rec), é desambiguado de acordo com o contexto da consulta (D_Q) e enriquecido semanticamente pela função (θ).

Seja $cobertura = \frac{|\{rec^*(t)\} \cap \{rec(t)\}|}{|\{\{rec^*(t)\}|}$.

Seja $ESA(a, Q)$ a função que calcula o valor médio da similaridade semântica para os elementos do conjunto $\theta_a(Q)$.

Assim sendo, as hipóteses para o RISO são:

- $cobertura_{RISO} \geq cobertura_{UBY}$
- $\mid \theta_{RISO}(Q) \mid \geq \mid \theta_{UBY}(Q) \mid$
- $ESA(\theta_{RISO}, Q) \geq ESA(\theta_{UBY}, Q)$

4.3 Enriquecimento Semântico de Conceitos

A seguir, é explicado em detalhes o funcionamento do módulo RISO-ES, o qual foi o objetivo deste trabalho de dissertação.

4.3.1 Extração de Conceitos

O processo de enriquecimento semântico conceitual inicia com a submissão de uma coleção de documentos que devem ser indexados seguindo a abordagem proposta pelo RISO-T. Neste processo, todos os documentos submetidos são previamente processados, gerando

como resultado, um vetor de termos extraídos de cada documento. Para realização desse procedimento, utiliza-se o mesmo processo (POS-Tagger) utilizado no módulo RISO-VTD, implementados e validados em [45].

A lista de vetores temáticos produzida para cada documento, contém sintagmas nominais e verbais que serão considerados conceitos de entrada para as fases posteriores. A Figura 4.6 apresenta um exemplo do processo que envolve a identificação de conceitos em texto não estruturado presentes nos documentos.

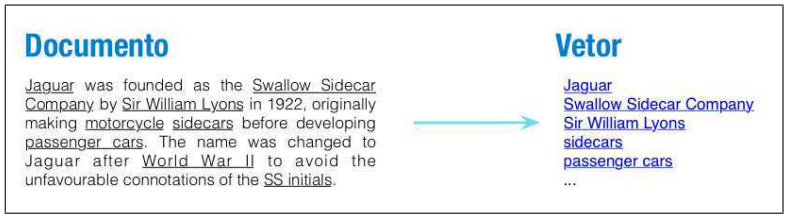


Figura 4.6: Processo de extração de conceitos de documentos textuais.

4.3.2 Enriquecimento Semântico

Para cada sintagma nominal encontrado ou deduzido pelo conjunto de regras da etapa de extração de conceitos, o processo de enriquecimento semântico tentará relacionar este elemento com pelo menos um conceito existente nas fontes externas e heterogêneas de informação. Este relacionamento poderá ocorrer de duas maneiras:

- *Perfeito*: Quando alguma das fontes de informação contém exatamente o sintagma extraído do documento textual(*string matching*), o processo de enriquecimento semântico ocorre sem a necessidade de nenhum processamento extra, apenas associando, o elemento extraído ao seu correspondente nas fontes externas. A Figura 4.7 apresenta um exemplo do procedimento para quando existe correspondência total entre o sintagma produzido e o conceito presente nas fontes de informação.

- *Parcial*: Por outro lado, quando nenhuma das fontes de informação contém exatamente o sintagma extraído do documento textual, o processo de enriquecimento semântico necessitará realizar um processamento extra com o objetivo de associar o elemento extraído ao seu correspondente mais próximo nas fontes externas.

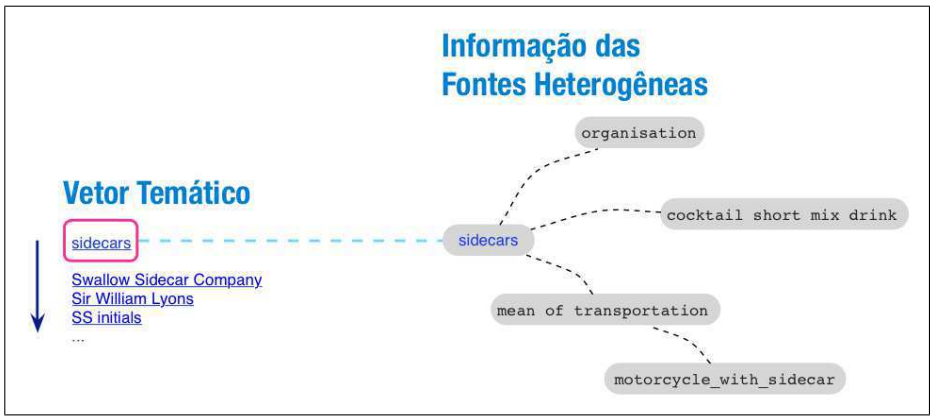


Figura 4.7: Processo de enriquecimento conceitual com relacionamento perfeito.

Primeiramente, se faz necessário formalizar a noção de *proximidade*. Neste caso, utiliza-se o *edit distance*, ou seja, o elemento mais *próximo* é aquele que possui a menor diferença na quantidade de termos quando comparado com o elemento original. Por exemplo, para o conceito *"Alice no país das maravilhas"* o conceito *"Alice"* diverge em 4 termos do primeiro conceito, por outro lado, o conceito *"Alice das maravilhas"* diverge apenas em 2 termos, sendo assim, *"Alice no país das maravilhas"* está mais próximo do conceito *"Alice das maravilhas"* do que de *"Alice"*.

Seguindo esta noção de proximidade (considerando as preposições), busca-se selecionar nas fontes externas de informação o conceito mais próximo para ser relacionado com o sintagma extraído do documento. Tal relacionamento, pode ocorrer de duas maneiras:

- *Generalização:* A generalização ocorre quando o sintagma proposto é relacionado com um conceito mais geral que ele. A Figura 4.8 apresenta o processo de relacionamento parcial associado a um conceito mais geral, onde Sir William Lyon é um William dentre todos os Willians existentes.
- *Especialização:* A especificação ocorre de maneira inversa à generalização, ou seja, quando o sintagma proposto é relacionado com um elemento mais específico que ele. Ainda utilizando o exemplo da Figura 4.8, caso o elemento mais próximo a Sir William Lyon, fosse por exemplo, Sir. William Lyon III, este por ser um elemento mais específico que Sir William Lyon, passaria a ser uma especificação

do sintagma proposto e a partir de então seria possível prosseguir com o processo de enriquecimento semântico.

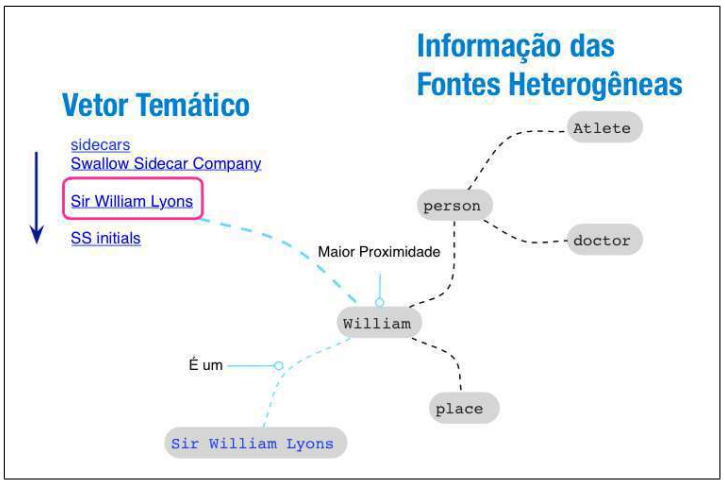


Figura 4.8: Processo de enriquecimento conceitual com relacionamento parcial.

No pior caso, onde não seja possível estabelecer nenhuma relação semântica, seja ela perfeita ou parcial, com as informações presentes nas fontes de informação, o conceito a ser enriquecido é associado diretamente ao elemento mais genérico possível, como por exemplo, *"Thing"* e, nestes casos, o enriquecimento semântico será incipiente em momento inicial, podendo ser melhorado ou por atualizações nas fontes de informação ou por estabelecimentos de relações de maneira manual. Vale salientar que, para o pior caso, o processo de indexação proporcionado posteriormente será equivalente ao convencional, ou seja, por palavra-chave.

Após estabelecido relacionamento com algum conceito das fontes de informação, inicia-se o processo de enriquecimento semântico. Nesta etapa, obtêm-se um conjunto de afirmações (*asserts*) que façam referência ao conceito submetido ao enriquecimento semântico. Por exemplo, para um sintagma *jaguar*, relacionado com um conceito *jaguar* nas fontes externas de informação é possível obter por meio de consulta as seguintes afirmações:

$$IsA(\underline{jaguar}, car)$$

$$IsA(\underline{jaguar}, automobile)$$

$$MemberOf(\underline{jaguar}, Phanteras)$$

$$IsA(\underline{jaguar}, \textit{mean of transportation})$$

$$PartOf(fangs, \underline{jaguar})$$

$$IsA(\underline{jaguar}, mammal)$$

$$IsA(\underline{jaguar}, animal)$$

$$IsA(\underline{jaguar}, vertebrate)$$

$$IsA(\underline{jaguar - xf}, jaguar)$$

$$IsA(\underline{jaguarf - type}, jaguar)$$

De posse do conjunto de afirmações relativas ao conceito, realiza-se um filtro sobre estas afirmações para serem eliminadas todas as que possuem a relação *TranslateOf* cuja semântica representa a tradução de um idioma para outro. Isso se faz necessário porque as fontes de informações possuem conceitos representados em vários idiomas, mas, utilizá-se neste trabalho apenas o idioma inglês, uma vez que a quantidade de termos de outros idiomas ainda é incipiente e considerá-las poderia interferir na qualidade dos resultados proporcionados pela abordagem.

4.3.3 Construção de Ontologia

De posse das afirmações resultantes das etapas anteriores, cada uma delas é analisada e organizada. Toda afirmação segue a seguinte estrutura de tripla proposta por Klyne [36]:

$$predicado(sujeito, objeto)$$

Inicialmente, o algoritmo proposto para indexação semântica de documentos proposto por este trabalho, receberá cada uma das afirmações, desmembrando-as ao mesmo tempo em que tentará relacionar sujeitos e objetos respeitando as relações semânticas de hierarquia e colateralidade (elementos que encontram-se em um mesmo nível em uma estrutura hierárquica), verificando a existência de novos relacionamentos entre afirmações e estabelecendo-os sempre que possível. A Figura 4.9 ilustra o procedimento de fusão e estruturação de duas afirmações. Neste exemplo, as fontes de informação possuíam o predicado $IsA(car, \textit{mean of transportation})$ e, assim, foi possível, a partir desta relação semântica,

estabelecer um relacionamento entre *"car"* e *"mean of transportation"* ampliando as informações sobre *"jaguar"*. .

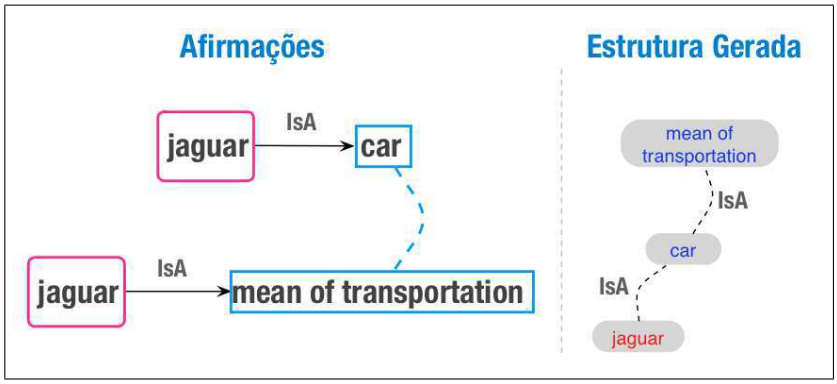


Figura 4.9: Processamento de triplas em momento inicial.

Por sua vez, a Figura 4.10 apresenta um exemplo mais avançado do processo de construção e estruturação das afirmações em mapas de tópicos semânticos. As linhas em cor

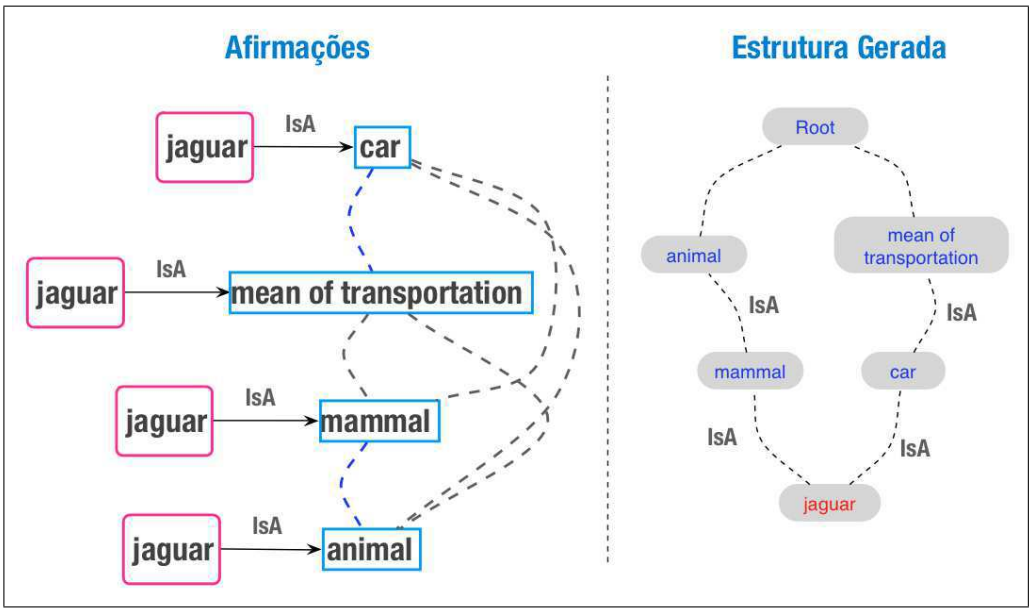


Figura 4.10: Processamento de triplas envolvendo afirmações em que não é possível estabelecer algumas relações.

azul apresentam conceitos que são possíveis relacionar por meio das fontes de informação e, por sua vez, as linhas em cor cinza indicam que as fontes de informação não apresentam relacionamento entre as entidades que tenta relacionar. Assim sendo, percebe-se que o

relacionamento entre *"mammal"* e *"mean of transportation"* não existe nas fontes de informação. Até este ponto, tem-se apenas uma rede semântica de informações, uma vez que, nesta etapa, inexiste o elemento *"ocorrência"* do mapa de tópicos, elemento este que só poderá ser atribuído, após ser realizado o processo de desambiguação.

Até esse momento, não é possível saber se o *"jaguar"* pertencente ao documento que se deseja indexar, se refere ao *animal* ou ao *veículo esportivo*. Para que seja possível gerar alguma conclusão sobre esse questionamento, se faz necessário uma etapa de desambiguação.

A Figura 4.11 apresenta a rede semântica resultante neste ponto do processamento. Percebe-se inicialmente que a estrutura formada é parcialmente completa, ou seja, ela preserva todas as relações ainda que sejam redundantes. Também é possível perceber que o elemento *jaguar*, sublinhado na figura, sofre sobrecarga de significado, pois possui relacionamento tanto com *mammal* quanto com *car*, ou seja, essa sobrecarga de significado representa uma ambiguidade que necessitará ser tratada antes de receber a indexação.

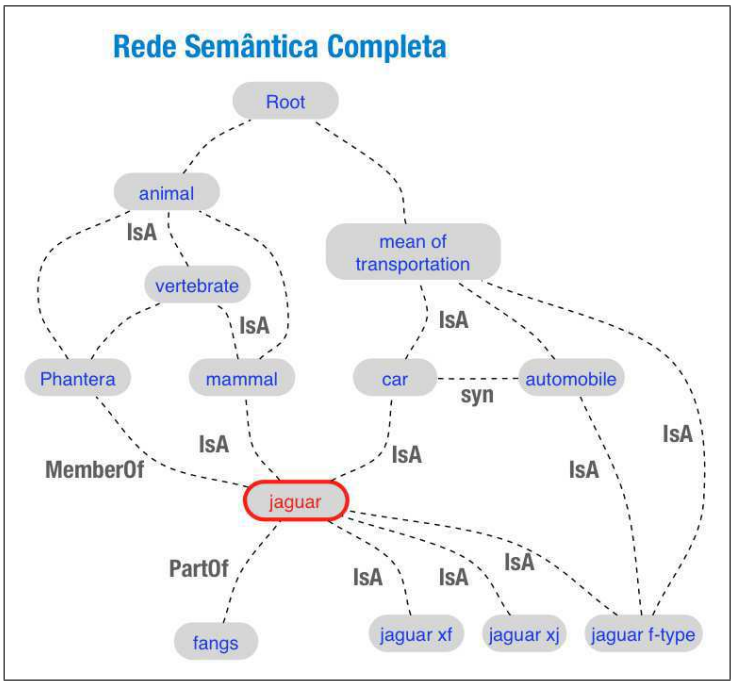


Figura 4.11: Enriquecimento conceitual com os relacionamentos presente nas fontes externas de informação.

Para a representação das afirmações utilizou-se a linguagem RDF e, para processar e

manipular essa linguagem, usou-se o framework JENA[4] adequado para manipulação de ontologias na linguagem Java. Cada tripla ou afirmação, logo após processada, foi submetida a este framework. A Figura 4.12 exemplifica parte da rede semântica apresentada graficamente na Figura 4.11, exibindo parte do arquivo de texto que contém a ontologia construída. Dentre as vantagens de organizar as informações em formato de ontologia estão: a capacidade de ser processada facilmente por computadores, visto que trata-se de um texto estruturado; e, a fundamentação matemática dessas estruturas que permitem explorar toda a potencialidade da máquina de inferência. Além disso, torna-se possível tanto a importação quanto exportação de informações, favorecendo a utilização do conhecimento produzido por qualquer ferramenta que dê suporte ao processamento de ontologias.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:coneitos="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rel="http://lsi.dsc.ufcg.edu.br/riso.owl#/r/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >

<rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/jaguar">
    <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/car"/>
    <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/animal"/>
    <rel:MemberOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Phantera"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
 </rdf:Description>

 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/jaguar f-type">
  <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/car"/>
  <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/automobile"/>
  <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/mean of transportation"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>

</rdf:RDF>
```

Figura 4.12: Estruturação de informações em ontologias utilizando RDF.

4.3.4 Desambiguação, Indexação Semântica e Construção de um Mapa de Tópicos Semântico

O processo de desambiguação inicia com a seleção de estruturas temáticas de domínios distintos que não se relacionam. Esses galhos se transformarão posteriormente em estruturas temáticas de domínios distintos, que serão fundamentais no processo de desambiguação. O algoritmo apresentado no código Fonte 4.1 mostra como são construídos os subgrafos de

[4]http://jena.apache.org/

domínios. A Figura 4.13 ilustra graficamente o que acontece após o processamento do algoritmo 4.1 de construção de subgrafos.

Código Fonte 4.1: Construção de subgrafos

```
List x = obtemFilhosDiretosDaRaiz(redeSemantica);
List subgrafos = [];

while(x.hasNext()){

        List filhos = obtemFilhosRecursivamente(x.next());
        List subgrafo = [];
        subgrafo.add(filhos);
        subgrafos.add(subgrafo);

}
```

Com a separação de conceitos em subgrafos de domínios distintos, criam-se os vetores temáticos que serão utilizados no processo de desambiguação. Cada *subgrafo* contém uma lista de termos que tratam de um mesmo contexto, e que serão utilizados neste processo.

Em paralelo a esse processo, transforma-se a rede semântica completa em uma rede semântica minimal, enxuta, onde toda regra que pode ser inferida deixa de ser afirmada. Esse procedimento torna-se extremamente útil para:

- *Armazenamento Eficiente de Afirmações:* Estruturar informações em ontologias, gerou a possibilidade de eliminar da estrutura semântica afirmações que podem ser inferidas por meio da máquina de inferência da própria ontologia e, assim, do ponto de vista de espaço em disco, disponibilizar mais espaço de armazenamento para novas afirmações. Para a grande quantidade de informação manipulada, tal atividade foi capaz de poupar terabytes[5] de alocação em disco.

- *Consultas e Ambiente de Interatividade:* Em um ambiente de consulta que utilizará a estrutura semântica criada para responder eficientemente aos usuários, utilizar um grafo completo proporciona opções redundantes no momento de desambiguação. Por

[5]Este é um valor aproximado obtido após comparar-se o tamanho em bytes de um arquivo contendo todas as afirmações obtidas após o processamento de um conjunto de documentos contra um arquivo contento as mesmas informações estruturado em ontologia. A lógica está, basicamente, na eliminação de informações repetidas e que podem ser inferidas por parte da ontologia.

exemplo, para um usuário que digita unicamente a palavra *jaguar* em um ambiente de busca que está diretamente conectado a uma estrutura semântica completa como apresentada na Figura 4.11, a tentativa de desambiguar o elemento consultado gerará opções redundantes. Assim, para a pergunta *"Qual elemento jaguar você se refere ?"* as opções seriam *"Phantera"*, *"Mammal"*, *"Car"*, *"Automobile"*, mas, percebe-se que, nesse caso, a redundância atrapalha a interação com o sistema, toda *"Phantera"* é *"Mammal"* e *"Car"* é sinônimo de *"Automobile"*, ou seja, a redundância acaba gerando um conjunto de opções repetidas e sem utilidade. Provavelmente este procedimento não atenderá a expectativa do usuário. Com a implementação da estrutura mínima, ou grafo minimal, apenas uma opção de cada item é exibida tornando a apresentação das opções mais limpa e a interação com o usuário mais útil e objetiva.

De posse dos vetores de domínio e com a estrutura semântica mínima, dá-se prosseguimento ao processo de desambiguação. No documento textual, extrai-se o parágrafo completo onde o conceito de interesse ocorre e realiza-se a comparação vetorial entre os termos do parágrafo e os elementos de cada um dos subgrafos com o objetivo de desambiguar o texto em questão. A Figura 4.14 apresenta o processo de desambiguação implementado pelo RISO-ES. Percebe-se que o texto comenta sobre *jaguar* no sentido de *veículo* e, em nenhum momento, houve correspondência entre os elementos do texto e os elementos do domínio *biologia*. Assim sendo, para este exemplo, a etapa de desambiguação afirmará que o conceito *jaguar* tratado no documento, refere-se ao *carro* e não a outras formas de interpretação. De posse dessa informação, já se torna possível a incorporação dos elementos *ocorrência* e *escopo* na estrutura de rede semântica, transformando-a em um mapa de tópicos semântico. A Figura 4.15 ilustra o estágio final da estrutura semântica e a indexação do documento, por sua vez, a Figura 4.16 apresenta computacionalmente como estas informações são representadas.

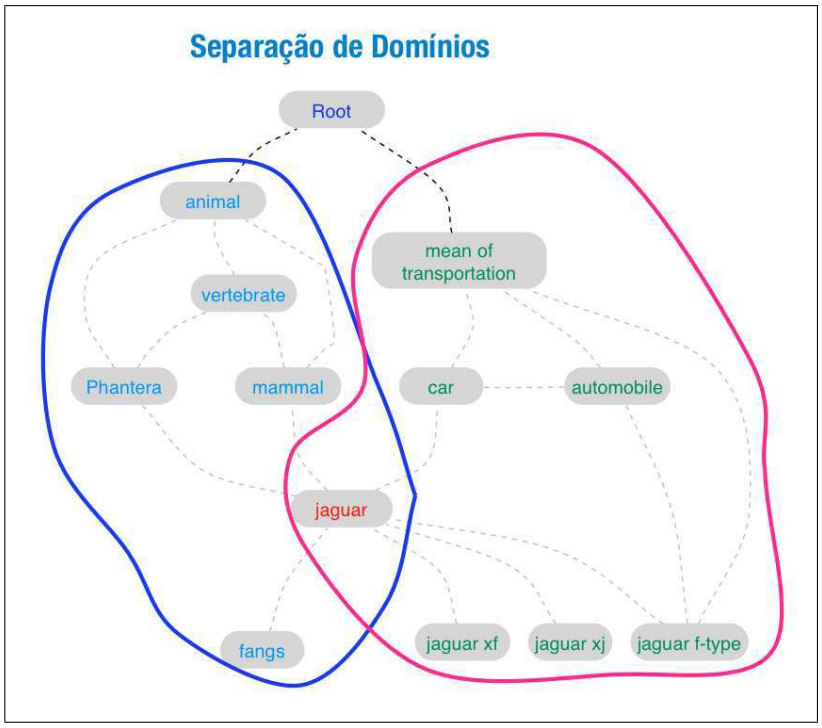


Figura 4.13: Construção dos conjuntos de elementos de domínios distintos.

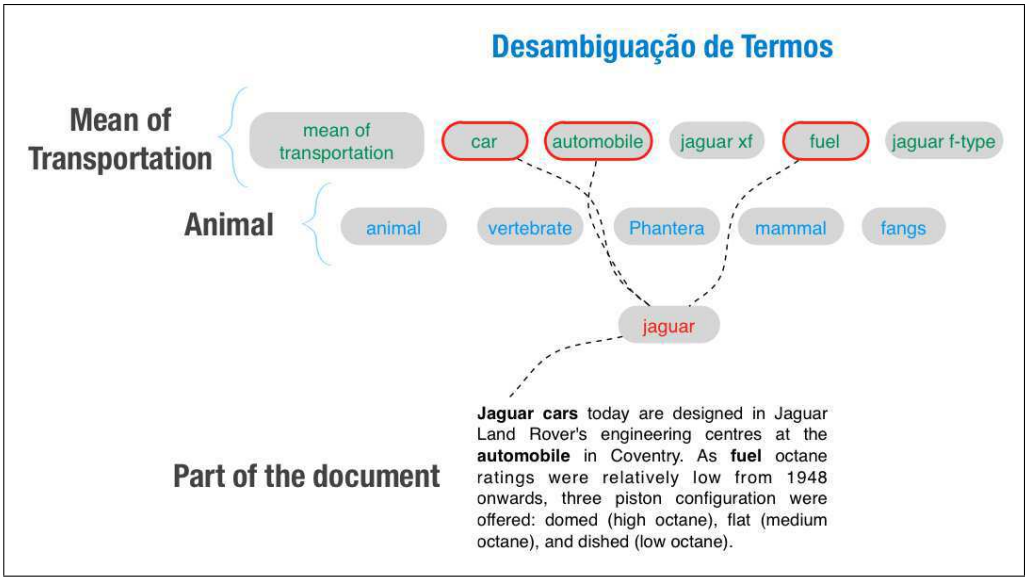


Figura 4.14: Processo de desambiguação implementado pelo RISO-ES.

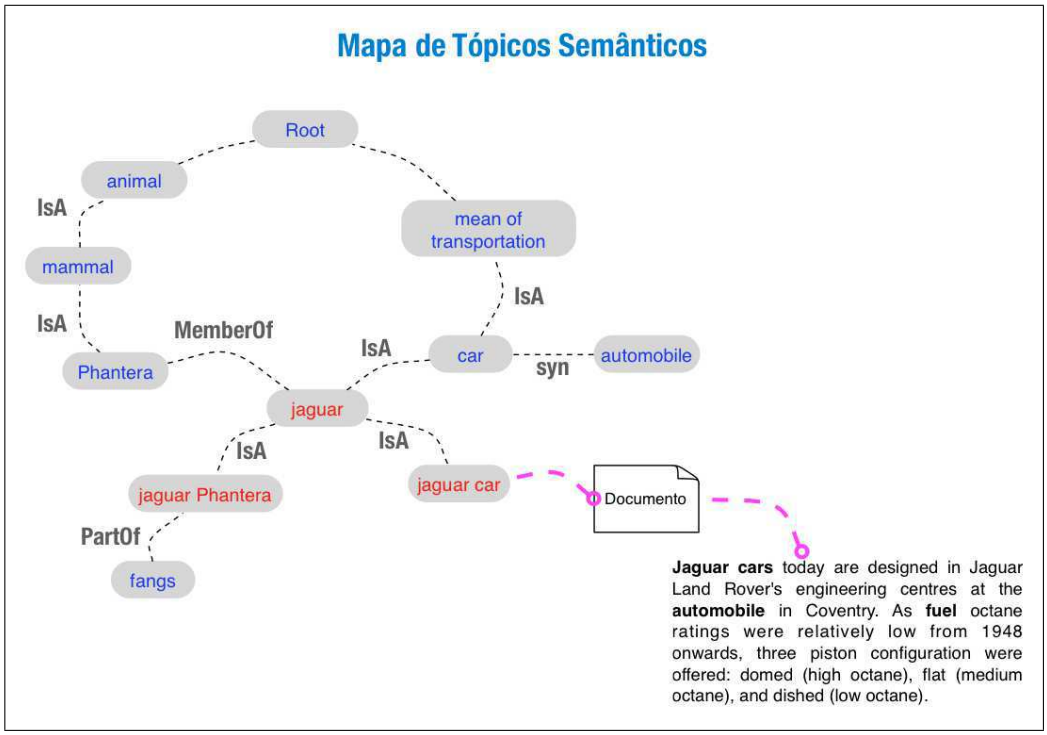


Figura 4.15: Grafo minimal e conexão com o documento.

```
466 <rdf:RDF
467     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
468     xmlns:coneitos="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/"
469     xmlns:owl="http://www.w3.org/2002/07/owl#"
470     xmlns:rel="http://lsi.dsc.ufcg.edu.br/riso.owl#/r/"
471     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
472     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
473
474 <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/jaguar">
475     <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/car"/>
476     <rel:MemberOf rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Phantera"/>
477     <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
478   </rdf:Description>
479
480   <rdf:Description rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Phantera">
481       <rel:IsA rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/animal"/>
482   </rdf:Description>
483
484   <rdf:Description rdf:about="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/jaguar {car}">
485     <rel:In rdf:resource="http://lsi.dsc.ufcg.edu.br/riso.owl#/c/en/Documento"/>
486     <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
487     </rdf:Description>
488
489 </rdf:RDF>
```

Figura 4.16: Grafo minimal e estrutura de tópicos semânticos representados computacionalmente.

4.3.5 Comentários Finais

A estrutura de mapa de tópicos produzida e estruturada em ontologia é armazenada em um banco de dados relacional, neste caso, utilizou-se o PostgresSQL[6], por possuir a característica de armazenar formatos de dados estruturados, processo este, próprio do framework Jena. A maneira como os dados foram estruturados, possibilita, sobretudo, consultas tanto sintáticas quanto semânticas por meio da linguagem SPARQL[7]. Caso o usuário deseje saber, por exemplo, todos os documentos que contenham a palavra *jaguar*, é possível desabilitar o motor de inferência e, assim, buscar por todos os documento com esta palavra (processo convencional). Se, por sua vez, o usuário desejar recuperar os documentos que tratam de *jaguar* no contexto *car* isso é possível localizando na estrutura o respectivo item *jaguar car* e recuperando todos os documentos relacionados a este item por meio da relação *rel:In* como apresentado na Figura 4.16, onde o tópico *jaguar car* contém os documentos que tratam de *jaguar* no contexto *car*. A implementação dessa abordagem reflete diretamente a formalização proposta na seção 4.2.

[6]Sistema gerenciador de banco de dados objeto relacional (SGBDOR), desenvolvido como projeto de código aberto. Para maiores informações: http://www.postgresql.org/.

[7]http://www.w3.org/TR/rdf-sparql-query/

Capítulo 5

Validação e Verificação

Este capítulo apresenta o design de experimentação utilizado para validar o RISO-ES, módulo do RISO-T, responsável pelo enriquecimento semântico proposto neste trabalho, como também, sua execução e discussão acerca dos resultados obtidos.

A validação apresentada comparou inicialmente ambas as propostas: RISO-T e UBY sob aspectos de cobertura e qualidade. No aspecto cobertura, buscou-se analisar quais das duas abordagens conseguia identificar a maior variedade de conceitos e, para cada conceito identificado, quais dos recursos apresentava maior número de relacionamentos (conexões). Para o aspecto qualidade, mediu-se o grau das relações semânticas entre o conceito base e outros conceitos associados a este por meio da estrutura semântica de cada abordagem.

Entende-se por grau das relações semânticas, a característica que mede a proximidade ou dependência que determinada relação exerce entre os conceitos que conecta. Por exemplo, conceitos unidos por meio de uma relação de *sinonímia* sejam mais fortemente relacionados que conceitos unidos por uma relação semântica do tipo *RelatedTo*, e assim sendo, o grau da relação semântica para o caso da relação de *sinoníma* é maior que a relação *RelatedTo*.

Valendo-se de um algoritmo que calcula a similaridade semântica (Semantic Similarity ou Semantic Relatedness) [69; 23] entre dois conceitos, foi possível calcular, de maneira objetiva, qual das duas abordagens é capaz de proporcionar um enriquecimento semântico composto por conceitos com maior grau de proximidade semântica e, consequentemente, mais úteis para o processo de indexação.

Toda execução do design experimental foi realizada em uma máquina macbook Pro com 4G de memória RAM e 1067 de DDR3, contendo um processador Intel Core 2 Duo de

2.4GHz e executando o sistema operacional Mac OS X em sua versão 10.6.8 com todas as atualizações aplicadas.

Para fins de reprodução e análise minuciosa de cada experimento, no apêndice A, encontram-se todos os detalhes necessários para realizar-se download dos softwares e bases de dados utilizadas para verificar e validar este trabalho, como também, o resultado completo de cada ensaio.

5.1 Cobertura e Conectividade

A análise de cobertura realizada nesta seção tem como finalidade comparar a eficiência de ambas as propostas (RISO-T versus UBY) mediante a capacidade que cada uma delas possui para detectar conceitos em suas fontes de informações e realizar o enriquecimento semântico. Nesta etapa, buscou-se submeter cada uma das redes semânticas de informações a uma gama de conceitos com grande variedade de características: nomes próprios, nomes de lugares, nomes de livros, datas (como por exemplo: *4 de julho*, *1945*), acrônimos de organizações e.g., *BB*, *OEA*, *MFI*, números e.g., *3.14*, *911*, siglas e definições coloquiais de algumas entidades.

Nesta etapa da análise, o único interesse foi o de medir a quantidade de elementos que cada proposta conseguiria identificar em sua estrutura semântica de informações e, uma vez detectado o conceito, medir a quantidade de conexões que este conceito possui com outros elementos da rede, quantificando, desta maneira, o grau de enriquecimento que ambas as abordagens proporcionam.

Obviamente, para um determinado conceito que possui um vasto número de conexões em uma dessas abordagens, não significa que o enriquecimento proporcionado é mais útil, uma vez que elementos inúteis e/ou com relações semânticas fracas podem preponderar. Assim sendo, avaliou-se a qualidade dessas conexões e do enriquecimento semântico proposto por ambas as propostas na seção 5.2.

5.1.1 Coleção de Teste Reuters-21578

A coleção de teste Reuters-21578[1] é amplamente difundida para validar trabalhos de Processamento de Linguagem Natural (PLN) em especial na área de Recuperação Categorizada de Texto [12; 15; 58; 13; 14], uma vez que, todos os dados da coleção foram coletados e etiquetados pelo Carnegie Group[2] e Reuters Ltd.[3].

Para esta validação, utilizou-se apenas as subcoleções de *nomes próprios*, *regiões geográficas*, *siglas e acrônimos* pelo fato de não apresentarem nenhuma marcação especial que necessitasse ser tratada (texto livre, diferente de *xml* e outros formatos de marcação). Cada uma dessas coleções de conceitos tenta capturar grande parte da variedade regional e cultural existente, apresentando vasta variedade de instâncias. Para coleção de nomes próprios, por exemplo, percebe-se a existência de elementos de origens diversas: árabes, nórdicas, latinas, saxônicas, entre outros. Para regiões geográficas, informações de países, cidades, municípios e até de ruas e vilarejos pouco conhecidos. No caso das siglas e acrônimos, o conjunto de teste explora desde instâncias conhecidas mundialmente, tais como: USA e EEUU até elementos pouco conhecidos como, por exemplo: IWS acrônimo para *International Wrestling Syndicate*.

A Figura 5.1 apresenta parte do processamento da coleção Reuters-21578. É possível perceber que os elementos da coleção recebem um valor entre colchetes referente ao número de conexões semânticas relacionadas a ele. Caso este valor seja zero, significa dizer que a abordagem submetida não foi capaz de identificar o elemento.

A Tabela 5.1 apresenta a porcentagem de conceitos identificados por cada uma das abordagens quando submetidas às subcoleções Reuters-21578. Em todos os casos, a rede semântica proposta pelo projeto RISO-T, por meio do módulo RISO-ES, apresentou maior cobertura que a estrutura proposta pelo projeto UBY, como é possível verificar graficamente na Figura 5.3. Por sua vez, a Tabela 5.2 apresenta uma comparação entre o número médio de conexões presentes entre o RISO-T e o UBY, considerando apenas conceitos identificados por ambas as abordagens. Para todas as coleções, o RISO-T também apresentou-se mais conectado que o UBY.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[2] http://www.thecarnegiegroup.com/

[3] http://www.reuters.com/

```
15 berge:[66]
16 bermúdez:[59]
17 beteta:[2]
18 blix:[8]
19 boesky:[1]
20 bond:[1219]
21 botha:[55]
22 bouey:[2]
23 braks:[0]
24 bresser-pereira:[0]
25 brodersohn:[0]
26 brundtland:[2]
27 camdessus:[2]
28 carlsson:[84]
29 caro:[118]
30 castelo-branco:[36]
31 castro:[699]
32 cavaco-silva:[2]
33 chaves:[94]
34 chen-muhua:[0]
35 chiang-ching-kuo:[8]
36 chien:[61]
37 chirac:[19]
38 ciampi:[5]
39 colombo:[485]
40 conable:[1]
41 concepcion:[52]
42 corrigan:[131]
43 cossiga:[5]
```

Figura 5.1: Exemplo de execução da coleção Reuters-21578.

Coleção Reuters(498)	RISO-T	UBY
Nomes Próprios(267)	**74.4%**	37.7%
Regiões Geográficas(175)	**98.2%**	80%
Siglas e Acrônimos(56)	**50%**	38.3%

Tabela 5.1: Comparação da cobertura entre UBY e RISO-T para a coleção Reuters-21578.

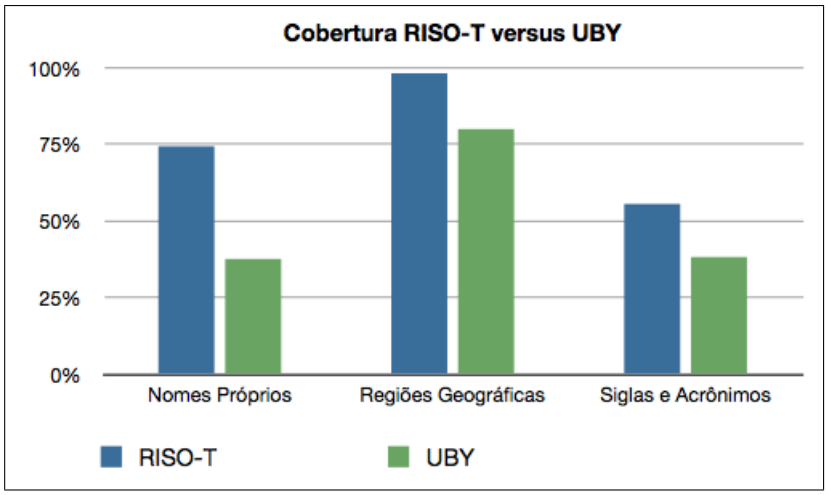


Figura 5.2: Cobertura do RISO-T versus UBY para coleção Reuters-21578.

Coleção Reuters	RISO-T	UBY
Nomes Próprios(267)	**165.2**	0.93
Regiões Geográficas(175)	**1185.2**	3.6
Siglas e Acrônimos(56)	**145.5**	0.8

Tabela 5.2: Número médio de conexões entre UBY e do RISO-T para a coleção Reuters-21578.

5.1.2 Analisadores Morfológicos

Analisadores Morfológicos são projetados para capturar grande variedade de conceitos, complexas formações de palavras compostas e codificação de caracteres mais amplas que ASCII de 8 bits. Em sua maioria, são utilizados como corretores ortográficos e possuem um conjunto de termos base que são previamente processados para serem posteriormente utilizados como sugestões.

Este experimento consistiu em utilizar conjuntos de termos de três analisadores morfológicos bastante utilizados: Hunspell[4], POS[5] e Agid-4[6]. O Hunspell, por exemplo, é utilizado por mais de 20 ferramentas amplamente difundidas no mercado, entre elas: *Apple's Mac OS X*, *Apache Solr*, *Eclipse*, *Google Chrome*, *LibreOffice*, *OpenOffice*, *Mozilla Firefox* e *Opera 10+*.

Buscou-se comparar as abordagens RISO-T, por meio do módulo RISO-ES, e UBY ainda sob os requisitos de cobertura e conectividade. Ainda que a coleção de testes Reuters-21578 seja amplamente utilizada em inúmeros trabalhos científicos como evidenciado, as subcoleções de testes úteis para validar esse trabalho apresentam um número muito restrito de instâncias. Visando suprir esta carência, realizou-se nova experimentação tendo como parâmetros de entrada todo conjunto de termos dos três analisadores morfológicos supracitados.

A Tabela 5.3 apresenta a porcentagem de conceitos identificados por cada uma das abordagens quando submetidas às subcoleções POS, Hunspell e Agid-4. Em todos os casos, a rede semântica proposta pelo projeto RISO-T, assim quando submetido à coleção Reuters-21578, também apresentou maior cobertura do que a estrutura proposta pelo projeto

[4]http://hunspell.sourceforge.net/

[5]http://sourceforge.net/projects/wordlist/files/POS/

[6]http://sourceforge.net/projects/wordlist/files/AGID/

UBY. Tal fato pode ser evidenciado graficamente na Figura 5.3. Por sua vez, a Tabela 5.4 apresenta uma comparação entre o número médio de conexões presentes entre o RISO-T e o UBY, considerando apenas conceitos identificados por ambas as abordagens. Para todas as coleções, evidenciou-se novamente que o RISO-T apresentou-se mais conectado que o UBY.

Analisadores Morfológicos(455.923)	RISO-T	UBY
POS(295.172)	**57.53%**	32.14%
Hunspell(48.246)	**92.07%**	76.94%
Agid-4(112.505)	**77.35%**	70.07%

Tabela 5.3: Comparação da Cobertura entre UBY e RISO-T para Analisadores Morfológicos.

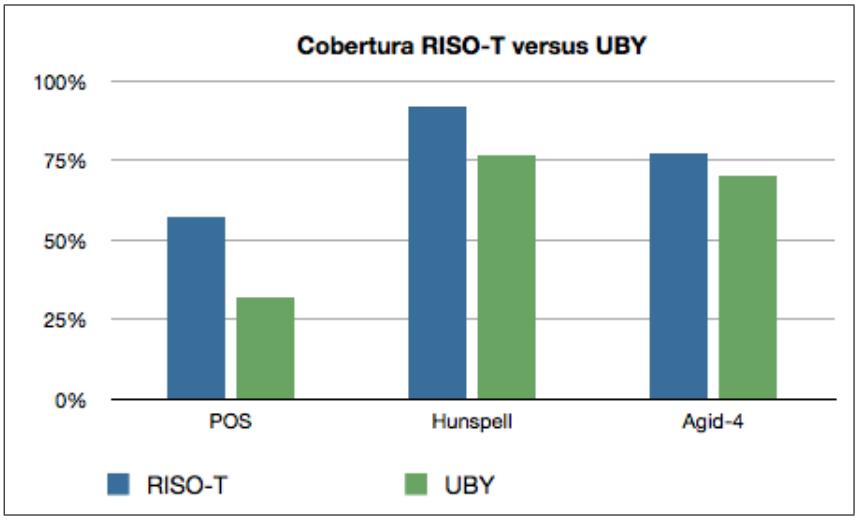


Figura 5.3: Cobertura RISO-T versus UBY para Analisadores Morfológicos.

5.2 Qualidade do Enriquecimento Semântico Proporcionado

Na seção 5.1, os únicos interesses foram os de avaliar o grau de cobertura e conectividade de conceitos submetidos a ambas abordagens (RISO-T e UBY) comparando-os sem, no entanto, analisar a qualidade dessa conexões. Nesta seção, serão discutidos aspectos relativos

Analisadores Morfológicos	RISO-T	UBY
POS	**598.08**	0.78
Hunspell	**2730.91**	2.45
Agid4	**1743.03**	1.94

Tabela 5.4: Conectividade do UBY e RISO-T quando submetidos às coleções de Analisadores Morfológicos.

à qualidade destas relações semânticas.

Uma vez que ambos os projetos têm por finalidade auxiliar sistemas de PLN em seu enriquecimento semântico, avaliou-se qual abordagem apresenta conexões semanticamente mais relevantes, e, consequentemente, consegue estabelecer relacionamentos mais úteis durante o processo de enriquecimento semântico. Por exemplo, para o conceito *"cachorro"* o sentido comum indica que conceitos, tais como: *"osso"*, *"pitbull"*, *"cão"* e *"pedigree"* são semanticamente mais relevantes que *"casinha"*, *"garagem"* e *"sapato"*. Isso porque: *"todo **cachorro** rói **osso**"*, *"**pitbull** é uma raça de **cachorro**"* e *"**cão** é um sinônimo de **cachorro**"*. Por sua vez, *"casinha"* não é algo propriamente de *"cachorro"*, ainda que *"cachorro"* possa morar em uma *"casinha"*, também pode morar em um *"canil"*, ou em outros ambientes. Além disso, *"casinha"*, também pode ser de *"boneca"* e/ou *"madeira"* e não genuinamente de *"cachorro"*. Outro exemplo é a relação existente entre *"cachorro"* e *"sapato"*. Apesar de geralmente *"cachorro"* gostar de morder *"sapato"*, ele também pode morder outros objetos e, além disso, *"sapato"* não é algo criado com a finalidade de ser mordido pelo *"cachorro"*, e sim, calçar os pés. Ou seja, *"sapato"* deve possuir uma relação semântica mais forte com o conceito *"pés"*, do que com o conceito *"cachorro"*. Por essas e outra gama de argumentações que estão fora do escopo desde trabalho, é possível afirmar que existe um conjunto de conceitos que estão mais relacionados com um determinado conceito base do que com outros. Nesta seção, buscou-se comparar qual das duas abordagens (RISO-T e UBY) é capaz de fornecer relações conceituais mais úteis no processo de enriquecimento semântico.

Para analisar a qualidade do enriquecimento proporcionado por ambas as abordagens, analisaram-se algoritmos da área de Similaridade Semântica (Semantic Similarity) [52], também conhecidos como Relacionamento Semântico (Semantic Relatedness). Trata-se de uma

área da Ciência da Computação que busca por meio de algoritmos, calcular o grau de semelhança semântica (significado e conteúdo) entre sentenças e documentos.

Para esta experimentação, utilizou-se o algoritmo estado da arte, largamente difundido pela comunidade científica para o cálculo da similaridade semântica, denominado ESA (Explicit Semantic Analysis).

5.2.1 Enriquecimento Semântico

Para medir a qualidade do enriquecimento semântico proporcionado por cada uma das abordagens, inicialmente foram selecionados aleatoriamente 2.400 conceitos de grafias distintas. Todos eles presentes em todas as coleções de testes utilizadas na etapa de validação da cobertura e conectividade na seção 5.1, realizando um total de 10 ensaios contendo 240 conceitos cada.

Sabendo-se que a rede semântica do projeto RISO-T apresentou uma superioridade de 26% no aspecto cobertura do que o projeto UBY, considerar os conceitos selecionados sem nenhum filtro, seria tendencioso e favoreceria o projeto RISO-T, uma vez que, muito provavelmente, uma taxa dos conceitos selecionados, seriam encontrados na rede semântica do RISO-T e não existiriam na rede semântica do projeto UBY. Dessa maneira, estes conceitos seriam contabilizados com o valor 0.0 durante o cálculo da qualidade do enriquecimento semântico proporcionado (*se não possui nenhuma informação para enriquecer um determinado conceito, tal contribuição para este conceito é zero*). Portanto, introduziu-se a restrição de que *os conceitos selecionados deveriam coexistir em ambas as abordagens* para que deficiências do aspecto cobertura não influenciasse a validação da qualidade do enriquecimento semântico.

Análises preliminares detectaram que conceitos com um número de conexões superior a 10.000 exigiriam muito tempo e poder computacional, tornando essa experimentação impraticável, caso o processamento desses dados fosse realizado de maneira centralizada. Além disso, grande parte dos conceitos presentes na rede semântica proposta pelo RISO-T, possui uma vasta quantidade de conexões, o que enfatizou a necessidade de se definir um limite superior de conexões. Por exemplo, o conceito *"big"* contém mais de 18.000 conexões, assim sendo, o esforço computacional necessário para processar conceitos com essa magnitude de conectividade seria muito grande. Com base no exposto, introduziu-se a res-

trição de que *conceitos selecionados não deveriam ultrapassar um grau de conectividade maior que 200*.

Inicialmente, submeteu-se cada um dos conceitos de cada ensaio, para serem enriquecidos por ambas as abordagens. Após a etapa de enriquecimento semântico, utilizou-se o algoritmo ESA para quantificar a qualidade das relações semânticas entre o conceito base enriquecido e os elementos oriundos do enriquecimento semântico proposto por cada uma das abordagens.

A Figura 5.6 exemplifica detalhes do processo de cálculos de qualidade fornecidos pelo algoritmo ESA para ambas as abordagens.

Apesar de que para muitos elementos das coleções, o resultado de ambas as abordagens se assemelhe e, além disso, o algoritmo ESA, utilizado para comparação de ambas as abordagens, estar apto tanto para fragmentar definições, quanto para associar termos, é importante enfatizar que o UBY e a abordagem aqui apresentada não fazem exatamente as mesmas atividades. Entretanto, o UBY é o mais próximo na literatura ao que é realizado por este trabalho.

Teste-t Pareado

Foram realizados 10 ensaios contendo cada um deles 240 conceitos selecionados aleatoriamente respeitando as duas restrições supracitadas na seção 5.2.1.

Cada conceito recebeu o enriquecimento semântico proposto por ambas as abordagens e ao final, foram calculadas as médias aparadas de cada uma delas, ou seja, a abordagem que apresentava o maior número de elementos enriquecidos, tinha a média calculada com apenas os n melhores elementos, onde n era a quantidade de elementos enriquecidos pela abordagem que apresentava o menor número de elementos.

Ao final de cada ensaio, foi calculada a média aritmética com os 240 valores proporcionados por cada uma das abordagens.

Por fim, realizou-se um *teste-t pareado* com as 10 médias gerais resultantes de cada ensaio, para verificar, se, estatisticamente, a qualidade dos resultados obtidos com o RISO-T eram superiores aos obtidos com o UBY. A Figura 5.4 apresenta a criação das estruturas vetoriais no sistema R^7 com os valores produzidos ao final de cada ensaio. A Figura 5.5 apre-

[7] http://www.leg.ufpr.br/ paulojus/embrapa/Rembrapa/

senta a execução do Teste-T tendo como hipóteses alternativas o fato de ambas as abordagens serem diferentes e da abordagem RISO-T ser superior a UBY. Em ambas as medições o *p-valor* foi inferior a 0.05, permitindo afirmar com 95% de confiança que a abordagem RISO-T apresentou em média 55% de superioridade quando comparada com o UBY.

```
> UBY<-c(0.124430106,0.122863293,0.117768275,0.121191789,
0.111348824,0.121278698,0.13776791,0.131231104,0.119236475,0.143224749)
> UBY
 [1] 0.1244301 0.1228633 0.1177683 0.1211918 0.1113488 0.1212787
 [7] 0.1377679 0.1312311 0.1192365 0.1432247
> RISO<-c
(0.669893006,0.6843255,0.652422822,0.686390962,0.67824159,0.677614606,0.71359336
0.696091794,0.671314871,0.707094933)
> RISO
 [1] 0.6698930 0.6843255 0.6524228 0.6863910 0.6782416 0.6776146
 [7] 0.7135934 0.6960918 0.6713149 0.7070949
> t.test(UBY,RISO,paired=T)
```

Figura 5.4: Construção dos vetores para o cálculo do Teste-t pareado.

```
> t.test(RISO,UBY,paired=T)

    Paired t-test

data:  RISO and UBY
t = 148.2667, df = 9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5501405 0.5671880
sample estimates:
mean of the differences
              0.5586642

> t.test(RISO,UBY,paired=T, alternative="greater")

    Paired t-test

data:  RISO and UBY
t = 148.2667, df = 9, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5517571       Inf
sample estimates:
mean of the differences
              0.5586642
```

Figura 5.5: Teste-t para ambas as hipóteses alternativas de igualdade e superioridade.

Análise Geral de Todos os Ensaios

Após o Teste-t apresentado, realizou-se uma análise geral com todos os conceitos enriquecidos ao longo dos 10 ensaios, medindo além da média dos valores atribuídos pelo algoritmo

ESA aos conceitos oriundos do enriquecimento semântico, o melhor valor apresentado pelo algoritmo ESA para o enriquecimento semântico em cada abordagem.

A Figura 5.7 apresenta graficamente as informações ordenadas dos melhores valores de enriquecimento proporcionados por cada uma das abordagens, quando submetidos ao algoritmo ESA. A abordagem RISO-T, por meio do módulo RISO-ES, forneceu o melhor valor de enriquecimento para 86% dos conceitos analisados contra 12% do UBY. Para 2% dos conceitos, tanto UBY, quanto RISO-T apresentaram os mesmos melhores valores e, assim sendo, contabilizou-se como empate.

A Figura 5.8 apresenta graficamente o cálculo das médias dos conceitos oriundos do enriquecimento semântico para cada um dos 2400 conceitos, quando submetidos as abordagens UBY e RISO-T.

A abordagem RISO-T apresentou média superior ao UBY em 64% dos conceitos semanticamente enriquecidos. Por sua vez, o projeto UBY foi superior à abordagem RISO-T para 31% dos conceitos calculados, tendo ambos, um percentual de empate de 5%.

Em ambas as análises de qualidade do enriquecimento semântico, a abordagem proposta pelo projeto RISO-T também demonstrou superioridade quando comparada com o UBY. A Tabela 5.5 apresenta o resumo do processo experimental para o requisito de qualidade.

	Melhor Valor	Melhor Média
UBY	12%	31%
RISO-T	**86%**	**64%**
Empate	2%	5%

Tabela 5.5: Quadro comparativo da validação da qualidade de enriquecimento semântico entre UBY e RISO-T.

5.3 Tempo de Processamento

Nesta seção, discute-se o requisito tempo de processamento. Para cada uma das experimentações apresentadas nas seções 5.1 e 5.2 mediu-se o tempo de processamento em minutos que cada uma das abordagens consumiu em média para produzir os resultados apresentados.

```
4)augury:
   (UBY)(2.3021645334333117E-6)-An omen or prediction; a foreboding; a prophecy.
   (UBY)(0.0)-A divination based on the appearance and behaviour of animals.
    Melhor UBY:2.3021645334333117E-6
   (RISO)(3.2208491424643163E-6)-work
   (RISO)(2.5393927116540294E-6)-inaugurate(be a precursor of)
   (RISO)(1.0)-us augury(A M-149)
   (RISO)(0.0)-war cloud(an ominous sign that war threatens)
   (RISO)(3.2300565838888085E-6)-write work
   (RISO)(0.13166514652061329)-augury(band)
   (RISO)(0.214585048725782)-augury of innocence(poems)
   (RISO)(0.0)-ship
   (RISO)(0.0)-fortune-telling
   (RISO)(2.93059224448877737E-9)-organisation
   (RISO)(0.0)-forebode
   (RISO)(1.0)-augury
   (RISO)(0.0)-mean of transportation
   (RISO)(2.8210039816264483E-6)-creative work
   (RISO)(0.0)-product
   (RISO)(2.1557123322674128E-6)-omen
   (RISO)(0.0)-organization
   (RISO)(5.964608317285131E-5)-augur
   (RISO)(8.689967403038707E-7)-band
   (RISO)(0.0)-music group
   (RISO)(0.1828027759009786)-augury(an event that is experienced as indicating important things to come)
   (RISO)(2.1464297460945695E-6)-omen(a sign of something about to happen)
```

Figura 5.6: Utilização do algoritmo ESA para quantificar o enriquecimento semântico proporcionados pelo UBY e RISO-T.

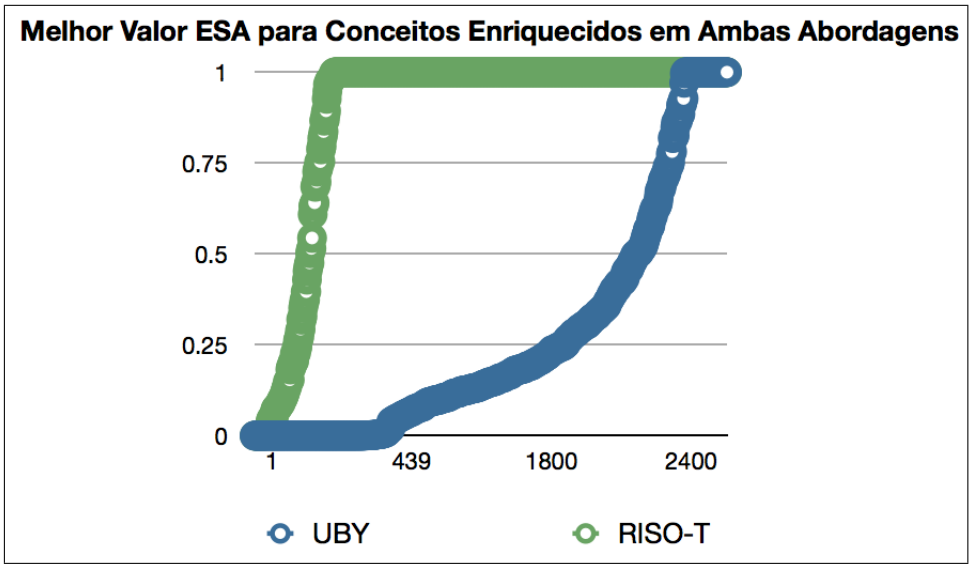


Figura 5.7: Melhor enriquecimento ponto a ponto.

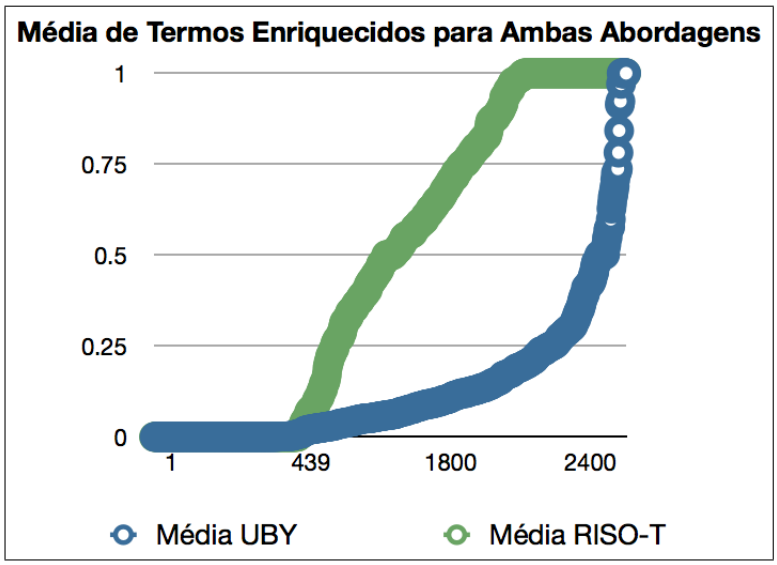


Figura 5.8: Melhor média entre os elementos propostos entre RISO-T e UBY.

5.3.1 Tempo de Processamento para o Cálculo da Cobertura e Conectividade

Na seção 5.1, ambas as ferramentas deveriam verificar a existência de determinado conceito em sua rede semântica e, em caso de detectada a existência, contar a quantidade de conexões que o conceito buscado possuía em sua estrutura semântica. As tabelas 5.6 e 5.7 apresentam o tempo gasto em minutos para realização do procedimento descrito por ambas as abordagens. Como evidenciado nas tabelas citadas e enfatizado pelos gráficos das figuras 5.9 e 5.10 o projeto UBY apresentou desempenho superior ao RISO-T no requisito tempo de processamento em todas as comparações para o cálculo de cobertura e conectividade.

5.3.2 Tempo de Processamento para o Cálculo do Enriquecimento Semântico

Na seção 5.2, ambas as ferramentas deveriam enriquecer semanticamente um determinado conceito, utilizando elementos de sua rede semântica. Enquanto o RISO-T necessitou de *240 minutos* para realizar o processamento, o projeto UBY utilizou apenas *42 minutos*. A Figura 5.11 apresenta graficamente a disparidade de tempo entre as duas abordagens. Para esta atividade, o UBY também apresentou um tempo de processamento inferior ao projeto

RISO-T.

5.4 Considerações Finais

Para todas as coleções de teste utilizadas, o projeto RISO-T apresentou uma cobertura média superior de 19.72%, bem como uma superioridade média da qualidade do enriquecimento semântico de 54%, quando comparado com o projeto UBY. Contudo, no requisito tempo de processamento, o projeto UBY foi mais eficiente, necessitando em média de apenas 28.14% do tempo necessário para que o projeto RISO-T execute as mesmas atividades.

Não foi encontrado na literatura nenhum outro trabalho que mesclasse fontes heterogêneas com características distintas (dicionários, enciclopédias e sentido comum) em um único ambiente e que usufruísse destas informações para enriquecer conceitos. Além disso, este trabalho ainda realiza a desambiguação de termos para a correta indexação conceitual de documentos textuais.

A utilização de analisadores morfológicos na etapa de validação permitiu verificar a qualidade deste trabalho para milhares de conceitos, diferente de outros trabalhos onde apenas dezenas de conceitos foram utilizados para validação ou empregada a técnica de *face validity*. Assim sendo, em ambos os aspectos o trabalho aqui apresentado inova e contribui com o estado da arte.

Analisadores Morfológicos	RISO-T	UBY
POS	390 *min*	**152 *min***
Hunspell	173 *min*	**31 *min***
Agid4	283 *min*	**53 *min***

Tabela 5.6: Tempo de processamento entre UBY e RISO-T quando submetidos aos Analisadores Morfológicos.

Reuters-21578	RISO-T	UBY
Nomes Próprios	38 *min*	**12 *min***
Regiões Geográficas	25 *min*	**8 *min***
Siglas e Acrônimos	4 *min*	**1 *min***

Tabela 5.7: Tempo de processamento entre UBY e RISO-T quando submetidos às subcoleções Reuters-21578.

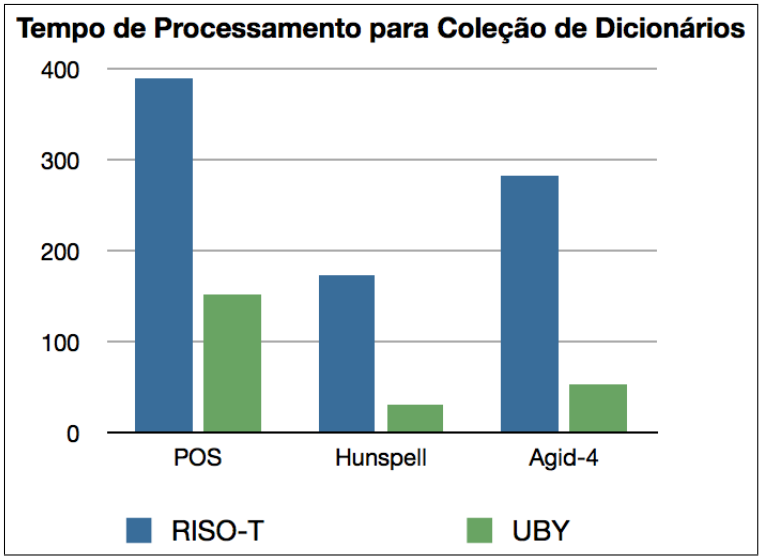


Figura 5.9: Tempo gasto em minutos para processar coleções de Analisadores Morfológicos entre RISO-T e UBY.

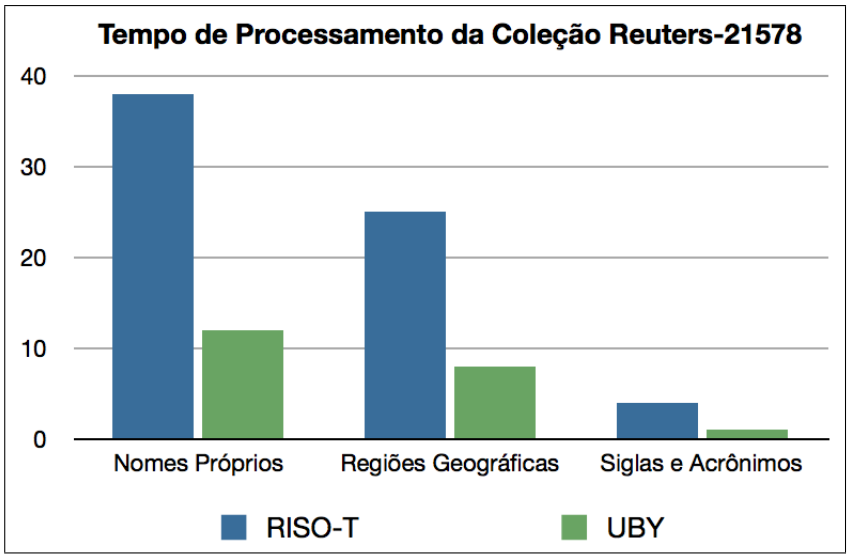


Figura 5.10: Tempo gasto em minutos para processar as subcoleções Reuters-21578 entre RISO-T e UBY.

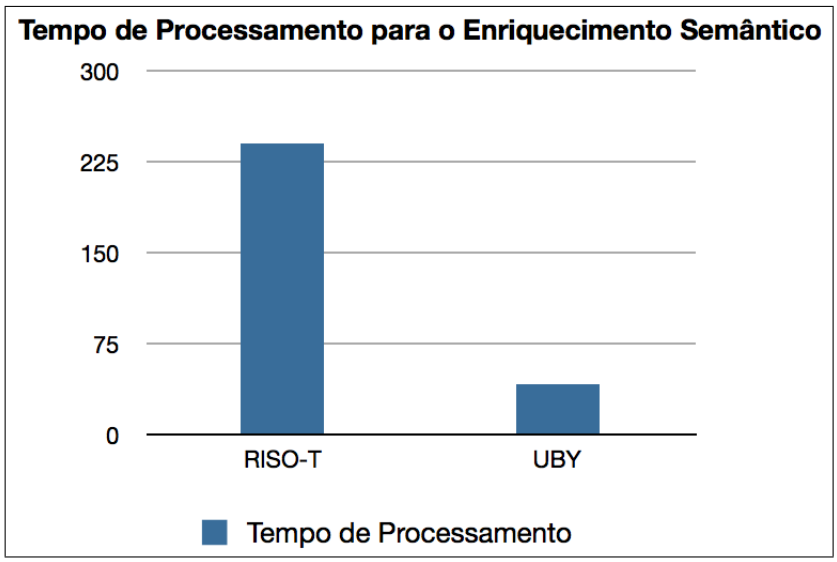


Figura 5.11: Tempo de processamento para calcular o enriquecimento semântico entre RISO-T e UBY.

Capítulo 6

Conclusão e Trabalhos Futuros

A proposta deste trabalho, consolidada por meio da construção do módulo de enriquecimento semântico do projeto RISO-T (RISO-ES) e validada no capítulo anterior, comprova melhores resultados que o projeto UBY nos requisitos de cobertura, conectividade e qualidades dos relacionamentos semânticos.

O casamento de fontes heterogêneas propostas pelo RISO-ES estabelece uma abordagem inovadora, unindo informações de dicionários, enciclopédias e opiniões de pessoas, incrementado a conjuntura convencional por meio do sentido comum humano. Tal abordagem, proporcionou estabelecer novas conexões, e construir um novo paradigma de indexação semântica estruturalmente mais abrangente. Estes elementos foram evidenciados quando comparados com um projeto semelhante, o UBY.

Apesar de necessitar de um tempo de processamento maior que o UBY para realização de todas as atividades propostas, este fato é parcialmente justificado pela necessidade de se processar uma quantidade maior de informações; obviamente, quanto maior o tamanho da entrada, maior o tempo necessário para realizar-se o processamento e, assim sendo, por possuir maior conectividade e mais elementos em sua estrutura semântica, o RISO-ES necessita de mais tempo para gerar seus resultados. Além disso, para se comunicar com as fontes externas, o RISO-ES necessita estabelecer conexões HTTP que geralmente são mais lentas que as conexões realizadas diretamente ao banco de dados, conforme realiza o UBY.

Para trabalhos futuros propõe-se:

- Implementar uma abordagem distribuída, paralelizando o processo de enriquecimento semântico, visando, assim, a realização da atividade em menor tempo.

- Estender a abordagem RISO-T para uma proposta multilíngue.
- Utilizar algoritmos estado da arte para identificação automática de conceitos, tais como, o DBpedia Spotlight[1] que consiste em utilizar as informações de artigos da Wikipédia na marcação de conceitos.
- Estender o arcabouço arquitetural proposto para o RISO-T para outras mídias: sons (RISO-A), imagens (RISO-I), vídeos (RISO-V) e, posteriormente, construir o projeto RISOM (Recuperação de Informação Semântica de Objetos Multimídia).
- Unir as informações de UBY e RISO-T e analisar se constituem abordagens complementares ou se UBY, em grande parte, constitui um subconjunto do RISO-T.
- Unir as informações de YAGO e RISO-T e analisar se constituem abordagens complementares ou se o YAGO, constitui um subconjunto do RISO-T.

[1] http://dbpedia-spotlight.github.com/demo/

Referências Bibliográficas

[1] Ahmed Abdelali, Jim Cowie, and Hamdy S Soliman. Improving query precision using semantic expansion. *Information processing & management*, 43(3):705–716, 2007.

[2] W. Abramowicz and H.C. Mayr. *Technologies for business information systems*. Springer, 2007.

[3] M. Agosti. *Information access through search engines and digital libraries*, volume 22. Springer, 2007.

[4] J. Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. Technical report, DTIC Document, 2005.

[5] N.J. Belkin. Some (what) grand challenges for information retrieval. In *ACM SIGIR Forum*, volume 42, pages 47–54. ACM, 2008.

[6] M.D. Bui and B.M. Duc. *Real-time object uniform design methodology with UML*. Springer, 2007.

[7] J. Callan, J. Allan, C.L.A. Clarke, S. Dumais, D.A. Evans, M. Sanderson, and C.X. Zhai. Meeting of the minds: an information retrieval research agenda. In *ACM SIGIR Forum*, volume 41, pages 25–34. ACM, 2007.

[8] M.L.A. Campos and H.E. Gomes. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. *Perspectivas em ciência da informação*, 11(3):348–359, 2006.

[9] C.R. Cavalcanti. *Indexação & tesauro; metodologia & técnicas: edição preliminar*. ABDF, 1978.

[10] C.A. Chahine, N. Chaignaud, J. Kotowicz, and J. Pecuchet. Conceptual indexing of documents using wikipedia. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 1, pages 195–202. IEEE, 2011.

[11] B. Chandrasekaran, J.R. Josephson, and V.R. Benjamins. What are ontologies, and why do we need them? *Intelligent Systems and Their Applications, IEEE*, 14(1):20–26, 1999.

[12] P.J. Cheng, M.Y. Kan, W. Lam, and P. Nakov. *Information Retrieval Technology: 6th Asia Information Retrieval Societies Conference, AIRS 2010, Taipei, Taiwan, December 1-3, 2010, Proceedings*, volume 6458. Springer, 2011.

[13] P. Cimiano. *Ontology learning and population from text: algorithms, evaluation and applications*, volume 27. Springer, 2006.

[14] P. Cimiano, A. Hotho, G. Stumme, and J. Tane. Conceptual knowledge processing with formal concept analysis and ontologies. *Concept Lattices*, pages 199–200, 2004.

[15] K. Coursey, R. Mihalcea, and W. Moen. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 210–218. Association for Computational Linguistics, 2009.

[16] I. Dahlberg. Teoria do conceito. *Ciência da informação*, 7(2), 1978.

[17] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[18] O. Egozi, E. Gabrilovich, and S. Markovitch. Concept-based feature generation and selection for information retrieval. *AAAIí08*, 2008.

[19] M.W. Evens. *Relational models of the lexicon: Representing knowledge in semantic networks*. Cambridge University Press, 2009.

[20] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[21] C. Fellbaum. *WordNet: an electronic lexical database*. Language, speech, and communication. MIT Press, 1998.

[22] O. Ferschke, T. Zesch, and I. Gurevych. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102. Citeseer, 2011.

[23] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007.

[24] L.M. Garshol. What are topic maps. In *XML. com*, 2002.

[25] L.M. Garshol. Metadata? thesauri? taxonomies? topic maps! making sense of it all. *Journal of information science*, 30(4):378–391, 2004.

[26] Salton Gerard and J McGILL Michael. Introduction to modern information retrieval, 1983.

[27] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[28] A. Goker and J. Davies. *Information retrieval: searching in the 21st century*. Wiley, 2009.

[29] T.R. Gruber et al. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 43(5):907–928, 1995.

[30] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human Computer Studies*, 43(5):625–640, 1995.

[31] Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter*

of the Association for Computational Linguistics (EACL 2012), pages 580–590, April 2012.

[32] E.H. Han and G. Karypis. Centroid-based document classification: Analysis and experimental results. *Principles of Data Mining and Knowledge Discovery*, pages 116–123, 2000.

[33] B. Harrington. A semantic network approach to measuring relatedness. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 356–364. Association for Computational Linguistics, 2010.

[34] B. He and I. Ounis. Studying query expansion effectiveness. *Advances in Information Retrieval*, pages 611–619, 2009.

[35] M. Horridge, H. Knublauch, A. Rector, R. Stevens, and C. Wroe. A practical guide to building owl ontologies using the protégé-owl plugin and co-ode tools edition 1.0. *University of Manchester*, 2004.

[36] G. Klyne, J.J. Carroll, and B. McBride. Resource description framework (rdf): Concepts and abstract syntax. *W3C recommendation*, 10, 2004.

[37] S. Kok and P. Domingos. Extracting semantic networks from text via relational clustering. *Machine Learning and Knowledge Discovery in Databases*, pages 624–639, 2008.

[38] G. Kowalski. *Information retrieval architecture and algorithms*. Springer, 2010.

[39] G. Kowalski. *Information Retrieval Architecture and Algorithms*. Springer, 2011.

[40] C.A. Kumar. Knowledge discovery in data using formal concept analysis and random projections. *International Journal of Applied Mathematics and Computer Science*, 21(4):745–756, 2011.

[41] R.H. Laan and G.I.S. Ferreira. Thesaurus e terminologia. In *Congresso Brasileiro de Biblioteconomia e Documentação,(19.: 2000: Porto Alegre, RS). Anais. Porto Alegre, 2000*, 2000.

[42] C.H. Li, W. Song, and S.C. Park. An automatically constructed thesaurus for neural network based document categorization. *Expert Systems With Applications*, 36(8):10969–10975, 2009.

[43] Y. Li, F.R. Reiss, and L. Chiticariu. Systemt: a declarative information extraction system. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 109–114. Association for Computational Linguistics, 2011.

[44] J. Lyons. Semântica vol. i e ii, 1979.

[45] C. Magna. *Criação de Vetores Temáticos de Domínios para a Desambiguação Polissêmica de Termos*. Number 1. UFCG, 2012.

[46] D. Milne. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*, pages 157–193, 2007.

[47] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann. Dbpedia and the live extraction of structured data from wikipedia. *Program: electronic library and information systems*, 46(2):157–181, 2012.

[48] C. Müller and I. Gurevych. Semantically enhanced term frequency. *Advances in Information Retrieval*, pages 598–601, 2010.

[49] V. Nastase and M. Strube. Decoding wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd national conference on Artificial intelligence*, volume 2, pages 1219–1224, 2008.

[50] N.F. Noy, D.L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[51] A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the web of concepts: Extracting concepts from large datasets. *Proceedings of the VLDB Endowment*, 3(1-2):566–577, 2010.

[52] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[53] P.M. Roget. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company, 1911.

[54] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 213–220. ACM, 2003.

[55] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[56] M. Scriven. *Evaluation thesaurus*. Sage Publications, Incorporated, 1991.

[57] W. Song, J. Yang, C. Li, and S. Park. Intelligent information retrieval system using automatic thesaurus construction. *International Journal of General Systems*, 40(04):395–415, 2011.

[58] S. Staab. *Handbook on ontologies*. Springer Verlag, 2009.

[59] R. Studer, S. Grimm, and A. Abecker. *Semantic web services: concepts, technologies, and applications*. Springer, 2007.

[60] F.M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.

[61] T. Tilley, R. Cole, P. Becker, and P. Eklund. A survey of formal concept analysis support for software engineering activities. *Formal Concept Analysis*, pages 250–271, 2005.

[62] J. Tramullas Saz. Propuestas de concepto y definición de la biblioteca digital. In *III Jornadas de Bibliotecas Digitales:(JBIDI'02): El Escorial (Madrid) 18-19 de Noviembre de 2002*, pages 11–20. Grupo de Ingeniería del Software, 2002.

[63] Princeton University. Wordnet, 2011.

[64] M. Uschold, M. Gruninger, et al. Ontologies: Principles, methods and applications. *Knowledge engineering review*, 11(2):93–136, 1996.

[65] J. Wales and L. Sanger. Wikipedia, 2011.

[66] J. Wang, Y. Wu, X. Liu, and X. Gao. Knowledge acquisition method from domain text based on theme logic model and artificial neural network. *Expert Systems with Applications*, 37(1):267–275, 2010.

[67] W. Weiszflog. Introdução: Novos termos. *MICHAELIS: moderno dicionário da língua portuguesa. São Paulo: Companhia Melhoramentos*, 1998.

[68] K. Weller. *Knowledge Representation in the Social Semantic Web*. Knowledge & Information: Studies in Information Science. De Gruyter Saur, 2010.

[69] I.H. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[70] P.R. Wojtinnek and S. Pulman. Semantic relatedness from automatically generated semantic networks. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS'11)*, 2011.

[71] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, volume 15, page 60, 2008.

[72] J. Zobel and A. Moffat. Exploring the similarity space. In *ACM SIGIR Forum*, volume 32, pages 18–34. ACM, 1998.

Apêndice A

Detalhes da Validação e Verificação

Com o objetivo de apresentar clareza sobre os dados utilizados na validação, fornecemos a possibilidade de download de todos os arquivos necessários para que o experimento possa ser reproduzido e analisado.

A.1 Enriquecimento Semântico

A.1.1 Software ESA

http://code.google.com/p/research-esa/

A.1.2 Conceitos Selecionados

A Figura A.1 apresenta parte dos conceitos selecionados para a etapa de validação da qualidade do enriquecimento semântico. O arquivo completo encontra-se disponível para download no seguinte endereço:

https://docs.google.com/file/d/0ByMErZ8giK2vX0VhZjdaTTZLQ2c/edit

Execução RISO-ES

A Figura A.2 apresenta parte do resultado da execução da abordagem RISO-ES ao conjunto de conceitos apresentados na seção A.1.2. O arquivo completo encontra-se disponível para download no seguinte endereço:

https://docs.google.com/file/d/0ByMErZ8giK2vN016ekNDMGhtQVk/edit

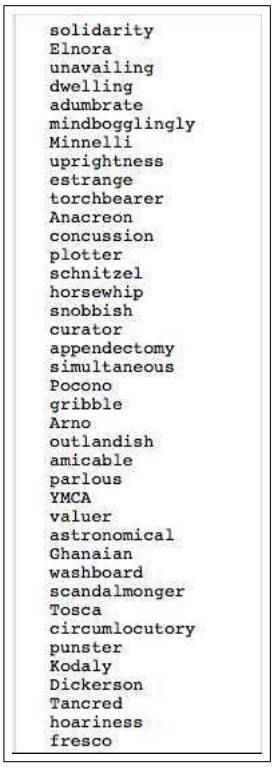


Figura A.1: Conceitos selecionados para o experimento.

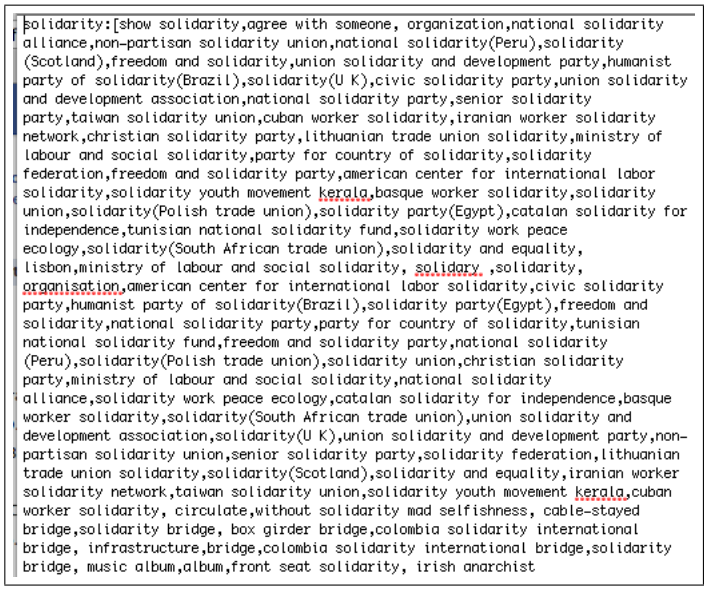


Figura A.2: Parte do processamento dos conceitos selecionados submetidos ao RISO-ES.

Execução UBY

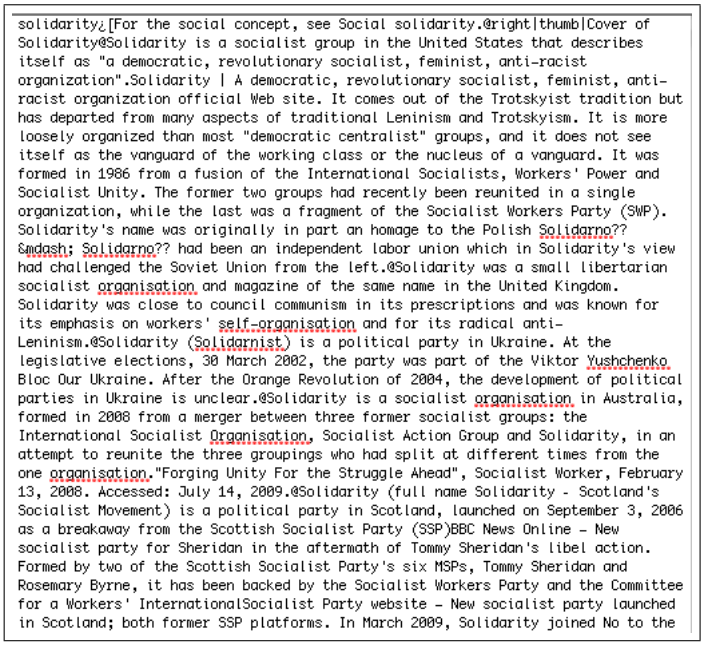


Figura A.3: Parte do processamento dos conceitos selecionados submetidos ao UBY.

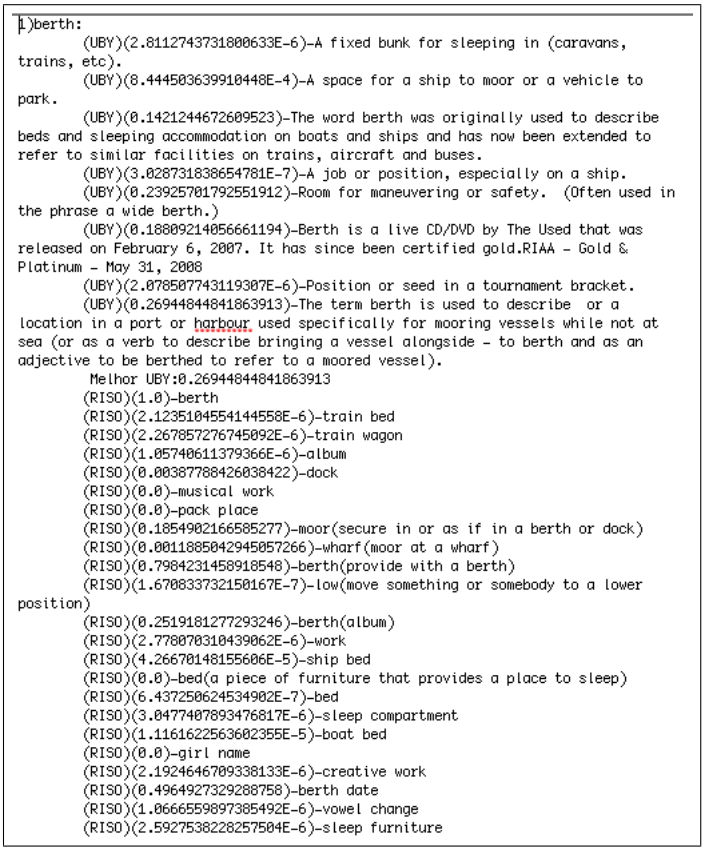
A Figura A.3 apresenta parte do resultado da execução da abordagem UBY ao conjunto de conceitos apresentados na seção A.1.2.O arquivo completo encontra-se disponível para download no seguinte endereço:

https://docs.google.com/file/d/0ByMErZ8giK2vZUtZc0ltdXdPM0U/edit

Resultado do Experimento

A Figura A.4 apresenta parte do resultado final da contabilização da qualidade entre as abordagens UBY e RISO-ES. O arquivo completo encontra-se disponível para download no seguinte endereço:

https://docs.google.com/file/d/0ByMErZ8giK2vWTd5R0hJMXJqSW8/edit



```
1)berth:
        (UBY)(2.8112743731800633E-6)-A fixed bunk for sleeping in (caravans,
trains, etc).
        (UBY)(8.444503639910448E-4)-A space for a ship to moor or a vehicle to
park.
        (UBY)(0.1421244672609523)-The word berth was originally used to describe
beds and sleeping accommodation on boats and ships and has now been extended to
refer to similar facilities on trains, aircraft and buses.
        (UBY)(3.028731838654781E-7)-A job or position, especially on a ship.
        (UBY)(0.2392570179255191 2)-Room for maneuvering or safety.  (Often used in
the phrase a wide berth.)
        (UBY)(0.18809214056661194)-Berth is a live CD/DVD by The Used that was
released on February 6, 2007. It has since been certified gold.RIAA - Gold &
Platinum - May 31, 2008
        (UBY)(2.078507743119307E-6)-Position or seed in a tournament bracket.
        (UBY)(0.2694484484186391 3)-The term berth is used to describe  or a
location in a port or harbour used specifically for mooring vessels while not at
sea (or as a verb to describe bringing a vessel alongside - to berth and as an
adjective to be berthed to refer to a moored vessel).
         Melhor UBY:0.26944844841863913
        (RISO)(1.0)-berth
        (RISO)(2.1235104554144558E-6)-train bed
        (RISO)(2.267857276745092E-6)-train wagon
        (RISO)(1.05740611379366E-6)-album
        (RISO)(0.00387788426038422)-dock
        (RISO)(0.0)-musical work
        (RISO)(0.0)-pack place
        (RISO)(0.1854902166585277)-moor(secure in or as if in a berth or dock)
        (RISO)(0.0011885042945057266)-wharf(moor at a wharf)
        (RISO)(0.7984231458918548)-berth(provide with a berth)
        (RISO)(1.670833732150167E-7)-low(move something or somebody to a lower
position)
        (RISO)(0.2519181277293246)-berth(album)
        (RISO)(2.778070310439062E-6)-work
        (RISO)(4.26670148155606E-5)-ship bed
        (RISO)(0.0)-bed(a piece of furniture that provides a place to sleep)
        (RISO)(6.437250624534902E-7)-bed
        (RISO)(3.0477407893476817E-6)-sleep compartment
        (RISO)(1.1161622563602355E-5)-boat bed
        (RISO)(0.0)-girl name
        (RISO)(2.1924646709338133E-6)-creative work
        (RISO)(0.4964927329288758)-berth date
        (RISO)(1.0666559897385492E-6)-vowel change
        (RISO)(2.5927538228257504E-6)-sleep furniture
```

Figura A.4: Parte da execução dos conceitos selecionados submetidos ao UBY.

A.2 Dados Reuters-21758

A.2.1 Siglas e Acrônimos

Execução do RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vZmhvaDBMdXRySjQ/edit

Execução do UBY

https://docs.google.com/file/d/0ByMErZ8giK2veXdPOF9XWXcxNHM/edit

A.2.2 Regiões Geográficas

Execução do RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vVE5zbHNxeFpKWm8/edit

Execução do UBY

https://docs.google.com/file/d/0ByMErZ8giK2vSFI1VkNyRUNLX1k/edit

A.2.3 Pessoas

Execução do RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vdHFFSHJULTFMVnM/edit

Execução do UBY

https://docs.google.com/file/d/0ByMErZ8giK2vN1pWMjloRGRnUWM/edit

A.3 Analisadores Morfológicos

A.3.1 POS

Execução RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vcVBYaW00R0hQaDQ/edit

Execução UBY

https://docs.google.com/file/d/0ByMErZ8giK2vMjFGVkRGYnlvY00/edit

A.3.2 Hunspell

Execução RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vRkMzQnhrdXZMWHc/edit

Execução UBY

https://docs.google.com/file/d/0ByMErZ8giK2vY3RNTUI2cWRGbVE/edit

A.3.3 Agid-4

Execução RISO-ES

https://docs.google.com/file/d/0ByMErZ8giK2vdWVlWnJPY1NvREk/edit

Execução UBY

https://docs.google.com/file/d/0ByMErZ8giK2vOG00ejdCaklFVEk/edit