IANNA MARIA SODRÉ FERREIRA DE SOUSA

# SISMULT – *SISTEMA DE INDEXAÇÃO SEMI-AUTOMÁTICA MULTILÍNGÜE*

Campina Grande - PB
Agosto de 1998

IANNA MARIA SODRÉ FERREIRA DE SOUSA

# SISMULT – *SISTEMA DE INDEXAÇÃO SEMI-AUTOMÁTICA MULTILÍNGÜE*

Dissertação apresentada ao Curso de Mestrado em Informática da Universidade Federal da Paraíba, em cumprimento parcial às exigências para obtenção do Grau de Mestre.

*Área de Concentração: Ciência da Computação*

*Sub-Área: Sistemas de Informação e Banco de Dados*

Orientador: Prof. Ulrich Schiel

**Campina Grande - PB**
**1998**


S725s Sousa, Ianna Maria Sodré Ferreira de.
SISMULT : Sistema de Indexação Semi-automática Multilingue / Ianna Maria Sodré Ferreira de Sousa. - Campina Grande, 1998.
76 f.

Dissertação (Mestrado em Informática) - Universidade Federal da Paraíba, Centro de Humanidades, 1998.
"Orientação : Prof. Dr. Ulrich Schiel".
Referências.

1. Sistema de Banco de Dados. 2. SISMULT - Sistema de Indexação Semi-Automática Multilíngue. 3. Sistema de Informação. 4. Dissertação - Informática. I. Schiel, Ulrich. II. Universidade Federal da Paraíba - Campina Grande (PB). III. Título

CDU 004.65(043)

SISMULT-SISTEMA DE INDEXAÇÃO SEMIAUTOMÁTICA MULTILÍNGUE

IANNA MARIA SODRÉ FERREIRA DE SOUSA

DISSERTAÇÃO APROVADA EM 31.08.1998

PROF. ULRICH SCHIEL, Dr.
Orientador

PROF. JACQUES PHILIPPE SAUVÉ, Ph.D
Examinador

PROFª HARUKA NAKAYAMA, Drª
Examinadora

CAMPINA GRANDE - PB

Aos Deuses,
protetores e cúmplices fiéis.

## AGRADECIMENTOS

Aos meus pais, pelo amor e pelos anos dedicados à minha educação.

Ao meu esposo e filho, pela compreensão nos momentos de ausência e pelo incentivo constante.

Ao prof. Ulrich, por nunca ter perdido a esperança.

Ao amigo Edberto pela troca de conhecimentos e por compartilhar as angústias e dúvidas.

A Rostand, pela colaboração na fase de implementação do trabalho.

E aos demais, que de alguma forma contribuíram na elaboração desta tese.

# ÍNDICE


1 **INTRODUÇÃO** ..... **1**

1.1 MOTIVAÇÃO ..... 4

1.2 REVISÃO BIBLIOGRÁFICA ..... 5

1.3 OBJETIVO ..... 10

1.4 ORGANIZAÇÃO DA DISSERTAÇÃO ..... 10

**2 FUNDAMENTAÇÃO TEÓRICA** ..... **12**

2.1 MÉTODO PARA EXTRAÇÃO DE TERMOS ..... 12

2.2 MÉTODO DA DECOMPOSIÇÃO RETANGULAR DE UMA RELAÇÃO BINÁRIA ..... 14

2.2.1 CONCEITOS BÁSICOS ..... 16

2.2.2 RELAÇÕES SEMÂNTICAS ..... 23

2.3 ALGORITMO DE PINTO ..... 26

2.4 THESAURUS MULTILÍNGÜE ..... 28

**3 MÉTODO PARA CONSTRUÇÃO SEMI-AUTOMÁTICA DE THESAURUS RETANGULAR MULTILÍNGÜE** ..... **31**

3.1 THESAURUS RETANGULAR MULTILÍNGÜE ..... 32

3.2 DESCRIÇÃO DO MÉTODO ..... 33

3.2.1 EXTRAÇÃO DE TERMOS ..... 35

3.2.1.1 Processando o Texto ..... 37

3.2.1.2 Palavras Compostas ..... 38

3.2.1.3 O Dicionário ..... 39

3.2.1.4 Identificando Conceitos Relevantes ..... 42

3.2.1.5 Matriz Conceito-Conceito e Cliques ..... 46

3.2.2 GERAÇÃO DOS RETÂNGULOS ÓTIMOS E ATUALIZAÇÃO DO THESAURUS CONCEITO-CONCEITO ..... 49

3.2.3 BASE DE DOCUMENTOS ..... 51

**4 PROTÓTIPO PARA CONSTRUÇÃO SEMI-AUTOMÁTICA DE THESAURUS RETANGULAR MULTILÍNGUE - SISMULT ..... 55**

4.1 INDEXAÇÃO DE DOCUMENTOS ..... 55

4.2 CONSULTAS ..... 63

4.3 EXPERIMENTAÇÃO ..... 66

**5 CONSIDERAÇÕES FINAIS ..... 68**

**REFERÊNCIAS BIBLIOGRÁFICAS ..... 70**

**APÊNDICE A – EXEMPLOS DE CLASSES CONCEITUAIS - DICIONÁRIO PRINCIPAL .75**

**APÊNDICE B – ESTRUTURA DE DADOS PARA GRAFO RETANGULAR ..... 76**

# LISTA DE FIGURAS

*Figura 2.1: Ilustração da noção de Relação Elementar* ..... 18

*Figura 2.2: Representações equivalentes de um mesmo retângulo* ..... 20

*Figura 2.3: Ilustração de Retângulo Ótimo* ..... 20

*Figura 2.4: Exemplo de relação genérica/especifica.* ..... 24

*Figura 2.5: Exemplo de sinônimos e pseudo-sinônimos* ..... 25

*Figura 2.6: Relação de vizinhança* ..... 26

*Figura 2.7: Representação de um Thesaurus Bilíngue* ..... 30

*Figura 3.1: Thesaurus Retangular Bilíngüe* ..... 33

*Figura 3.2: Etapas para a Construção de um Thesaurus* ..... 34

*Figura 3.3: Algoritmo simplificado do processo de indexação semi-automático.* ..... 36

*Figura 3.4 – Representação gráfica da estrutura de um dicionário* ..... 41

*Figura 3.5 – Grafo das distâncias entre as categorias gramaticais para um idioma* ..... 44

*Figura 3.6: Exemplos de Cliques* ..... 47

*Figura 3.7: Matriz conceito-conceito representando a Medida M entre os pares de conceitos* ..... 47

*Figura 3.8: Matriz binária conceito- conceito* ..... 48

*Figura 3.9: Grafo da matriz binária conceito-conceito* ..... 48

*Figura 3.10: Cliques extraídos* ..... 48

*Figura 3.11: Grafo de retângulos ótimos, em português, gerado a partir da matriz conceito-conceito.* ..... 50

*Figura 3.12: Matriz binária conceito-documento para um conjunto de quatro documentos* ..... 52

*Figura 4.14: Organização hierárquica (simplificada) da Base de Documentos para a matriz da*

*Figura 3.13: Grafo de retângulos ótimos para a base de documentos* ..... 53

*Figura 3.14: Grafo de retângulos ótimos simplificado em português* ..... 54

*Figura 4.1: Janela de gerenciamento do thesaurus.* ..... 57

*Figura 4.2: Janela de extração de termos* ..... 58

*Figura 4.3: Janela de Resolução de Conflitos* ..... 59

*Figura 4.4: Janela para inserção de novos termos ao dicionário* ..... 60

*Figura 4.5: Janela para conjugação de verbos* ..... 60

*Figura 4.6: Janela para inserção das flexões de termos substantivos e adjetivos.* ..... *61*

*Figura 4.7: Janela para atualização de stopwords.* ..... *62*

*Figura 4.8: Status do processamento do thesaurus* ..... *62*

*Figura 4.9: Tela de consulta do SISMULT* ..... *66*

# LISTA DE TABELAS

*Tabela 3.1: Exemplo do cálculo da Força de Ligação para um documento sobre orientação a objeto*.....46

*Tabela 4.1 – Conjunto de documentos processado pelo SISMULT*.....67

*Tabela 4.2 – Conjunto de documentos processado pelo SISMULT*.....67

# RESUMO

Com a difusão das bibliotecas digitais e da Internet, mais e mais textos em meio eletrônico, em diversos idiomas, se tornam acessíveis para um público amplo e geograficamente disperso. Isto torna necessário o desenvolvimento de ferramentas adequadas para facilitar a indexação, o armazenamento e a recuperação adequada de documentos referentes à informação pesquisada.

Este trabalho tem como objetivo apresentar um método para construção semi-automática de um thesaurus retangular multilíngüe, a partir de documentos eletrônicos, que auxiliará no processo de recuperação da informação, independente do idioma.

O método consiste em extrair termos semi-automaticamente do conjunto de documentos e utilizar a análise da co-ocorrência de termos para selecionar os termos relevantes, após consultar os dicionários unilíngües para determinar os termos abstratos. Os conceitos relevantes extraídos dos documentos são então representados por uma relação binária sobre a qual aplica-se o Método de Decomposição Retangular de uma Relação Binária para a obtenção dos retângulos que geram o thesaurus a partir de um algoritmo incremental.

Dicionários especiais e interações com o usuário são utilizados para determinar o contexto adequado para palavras ambíguas, além de eliminar flexões e associar um conceito abstrato para cada palavra.

O protótipo desenvolvido permite uma atualização contínua dos thesauri existentes com novos documentos, em diversos idiomas, e a realização de consultas multilíngües, além de permitir o acréscimo de novos idiomas.

# ABSTRACT

With the outspreed of the digital libraries and the Internet more and more electronic texts, written in several languages, become available for a wide and geographically dispersed public. This turns it's necessary to develop tools that facilitates indexing, representation and retrieval of multilingual documents.

This thesis presents a method for semiautomatic construction of a multilingual thesaurus, based on the indexing of electronic documents, in order to support a adequate information retrieval, independent of the language of the documents.

The method consists in extracting the terms of a document and to use an analysis of the co-occurrence of terms in order to determine its relevance. Using special unilingual dictionaries, abstract, language-independent terms are determined. Relevant concepts are represented as binary relations and, using the method of rectangular decomposition of Gammoudi, rectangles of pairs concept/document are determined and added to the existing thesaurus incrementally.

Special dictionaries and an interaction with the user determines the correct contexts for ambiguous terms, further on eliminating flexions and determining the abstract concepts.

A prototype has been developed which allows a continuous update of the existing thesaurus, indexing new documents, in several languages. It also supports multilingual queries and the addition of the new languages.

# 1 INTRODUÇÃO

O crescente volume de informação online, em vários idiomas, gerado e disseminado através de redes de computadores tem exigido de pesquisadores e profissionais a disponibilização aos usuários de recursos que auxiliem o armazenamento e recuperação de informações relevantes. A expectativa para o ano 2000 é de se pesquisar na Internet sobre um bilhão de páginas de dados [Kowalski, 1997].

Para informações estruturadas os Sistemas Gerenciadores de Bancos de Dados (SGBDs) têm sido usado com sucesso. Contudo, para dados textuais não estruturados, o gerenciamento, processamento e recuperação da informação ainda são tarefas complexas e problemáticas.

A linguagem escrita, contida em livros, artigos e documentos online, foi a primeira forma de armazenamento e fonte transmissora do conhecimento, universal, depois da fala. O problema é que a linguagem natural é prolixa e muitas palavras existem no texto apenas para compor a estrutura gramatical, num determinado idioma. Além do mais, o assunto de interesse pode estar espalhado em diversos textos diferentes, o que demanda muito tempo de leitura.

A recuperação eficaz de informações depende da análise correta e consistente do conteúdo do material a ser pesquisado e recuperado. A *indexação,* que consiste na extração de termos relevantes de um documento, é um método utilizado nesta análise e tem como objetivo determinar uma representação intermediária entre as informações textuais existentes e as consultas realizadas pelos usuários, a fim de

facilitar a pesquisa sobre o conteúdo das informações armazenadas ([Ferneda, 1997], [Frakes, 1992]).

Uma das formas de organização dos termos obtidos na indexação mais conhecidas e utilizadas nos sistemas de recuperação de informações é o *Thesaurus*. Um sistema de pesquisa pode automaticamente expandir ou restringir uma consulta do usuário, a fim de obter resultados mais precisos, usando a estrutura do thesaurus.

A palavra *Thesaurus* (derivada do grego para *depósito de riquezas)* ficou conhecida em meados do século XIX quando o Dr. Peter Mark Roget publicou o seu dicionário analógico **"Thesaurus of English Words and Phrases"**. Este dicionário tinha como destaque a disposição das palavras. As palavras não eram apresentadas em ordem alfabética, como nos demais dicionários, mas de acordo com as idéias que elas exprimiam.

Os bibliotecários foram os primeiros a usar um thesaurus para indexar manualmente os documentos ou livros de forma a encontrar rapidamente uma obra que diz respeito a um determinado assunto. Depois o thesaurus passou a ser usado como um elemento normalizador, permitindo ao usuário recuperar documentos por palavras chaves. Para indexar os documentos os especialistas usavam uma lista de termos proposta pelo thesaurus.

Um thesaurus constitui-se de um conjunto de termos, em linguagem natural, e um conjunto de relações semânticas entre estes termos, formando uma rede [Ferneda, 1997]. Além da relação hierárquica (termo genérico-termo específico) e a relação de equivalência (termos sinônimos), os thesauri incorporam uma relação associativa ou não hierárquica que refere-se aos termos relacionados, importantes na expansão da consulta .

Os thesauri podem ser apresentados ou visualizados de forma alfabética ou hierárquica. Na forma hierárquica, os termos são agrupados segundo sua posição na hierarquia (termo geral seguido de seus termos específicos, e assim por diante). Dentro de cada nível os termos podem ser agrupados logicamente ou alfabeticamente. Na forma alfabética, a lista de termos é apresentada alfabeticamente, incluindo em geral, notas de escopo e termos relacionados ([Aitchison, 1979], [Guidelines, 1997]).

Pelo fato da construção de thesauri ser altamente conceitual e de processo intensivo de conhecimento, os thesauri são construídos por especialistas no assunto, aplicando um processo geral de aquisição de conhecimento. Este processo inclui a coleção de fontes de conhecimento, identificação dos termos do thesaurus, e estabelecimento dos relacionamentos entre os termos do thesaurus. Destes passos, os dois últimos são de intensivo conhecimento. O processo inteiro pode consumir muito tempo de processamento.

Devido à complexidade da construção manual de thesauri, a construção automática se tornou um objetivo altamente desejável, tendo sido desenvolvidos vários trabalhos de pesquisa nos últimos anos ([Crouch, 1990], [Chen, 1993], [Chen, 1995], [Yuan, 1997], [Multites,1997]). As estratégias mais conhecidas empregam tecnologias estatísticas, lingüísticas ou mesmo de inteligência artificial, como redes neurais, para extrair informações conceituais e relacionais de uma grande base textual, e depois construir o thesaurus baseado na informação.

Como a construção de thesauri é complexa, eles não estão comumente disponíveis em muitos domínios. Ao contrário dos thesauri, outras formas de descrição de conhecimento, como os dicionários, estão largamente

disponíveis em muitos domínios. Comparados aos thesauri os dicionários, apesar de muito bem organizados, enfatizam as definições e o uso dos termos (informações conceitual e relacional).

## 1.1 MOTIVAÇÃO

"A habilidade de manipular publicações em muitos idiomas e acomodar vários idiomas no projeto de serviços de informação torna-se mais importante quando sistemas verdadeiramente internacionais se tornam operacionais" [Lancaster, 1986].

O thesaurus é bastante utilizado como estrutura de indexação e recuperação de informações. Foi introduzido com a necessidade de gerenciar grandes volumes de dados. No entanto, no decorrer dos anos sua construção tem permanecido manual, pelo fato de ser altamente conceitual e de processo intensivo de conhecimento, cujo custo é proibitivo. A construção manual é viável apenas quando o vocabulário é limitado, existe um especialista ou grupo motivado. Outrossim, especialistas diferentes podem construir thesauri diferentes a partir de uma mesma lista de palavras chaves.

Quanto à recuperação de informação, as pesquisas voltam-se para prover acesso efetivo e eficiente aos bancos de dados *unilíngües*, em geral em inglês, por ser o idioma padrão. Quando um usuário deseja recuperar documentos em um idioma qualquer, espera-se que o mesmo saiba formular uma consulta naquele

idioma. Contudo, existem algumas necessidades que não podem ser satisfeitas pelos sistemas de recuperação unilíngües, como por exemplo:

- numa coleção de documentos em vários idiomas seria necessário formular uma consulta em cada idioma;
- o usuário não é suficientemente fluente no idioma da coleção de documentos para expressar uma consulta naquele idioma, mas é capaz de fazer uso dos documentos recuperados;
- uma consulta formulada em um idioma não recupera documentos relacionados que estão em outro idioma.

A recuperação de informação multilíngüe é a recuperação de documentos baseada em consultas formuladas por um humano usando linguagem natural, independente do idioma em que os documentos e a consulta estão expressos. É a habilidade de formular uma consulta em um idioma e receber um documento em outro que distingue recuperação textual multilíngüe de recuperação textual unilíngüe [Oard, 1997a].

## 1.2 REVISÃO BIBLIOGRÁFICA

Em ciência da informação, o uso de um thesaurus ou base de conhecimento para recuperação de informação "inteligente" tem atraído a atenção de pesquisadores nos últimos anos.

Muitas das bases de conhecimento são geradas manualmente ou a partir de domínios específicos, usando o processo de aquisição de conhecimento, ou derivado de thesauri existentes. Por exemplo, CoalSORT [Monarch, 1987], uma interface baseada em conhecimento, facilita o uso de bancos de dados bibliográficos na tecnologia Coal, usando uma rede semântica. O sistema de linguagem médica unificado de Medicina (UMLS – Unified Medical Language System) da biblioteca do congresso americano tem como objetivo construir um sistema inteligente, automático, para a organização e recuperação da informação [Lindberg, 1990] [McGray, 1990]. O UMLS inclui um metathesaurus, uma rede semântica, e um mapa de fontes de informação. O metathesaurus contém informação sobre conceitos biomédicos e suas representações em mais de 10 vocabulários e thesauri diferentes.

Desde o princípio, a necessidade de se criar uma padronização para prevenir a criação de muitos vocabulários de assuntos indexados incongruentes e divergentes era evidente [Oard & Dorr, 1996]. As normas ISO 2788 e a ISO 5964, aprovadas em 1974 e 1978, descrevem como o conhecimento de um domínio pode ser incorporado em um thesaurus e identifica técnicas alternativas para o desenvolvimento de thesaurus, unilíngüe e multilíngüe, respectivamente. Em 1985 a norma ISO 5964 foi modificada pelo Comitê Internacional para a História da Arte (CIHAP) que imaginou um thesaurus multilíngüe para a terminologia da Arte e Arquitetura (AAT) incluindo cinco idiomas: inglês, francês, italiano, espanhol e alemão [Getty, 1997].

O EUROVOC do Parlamento Europeu é um exemplo de um moderno thesaurus multilíngüe segundo a norma ISO 5964. Primeiro publicado em 1984, EUROVOC agora inclui todos os nove idiomas oficiais da Comunidade Européia, e

porções dele têm sido traduzidas em outros idiomas. O projeto de thesaurus permanece caro, e este fato tem limitado os domínios em que a recuperação de vocabulário controlado tem sido aplicado. Mas o EUROVOC demonstra que uma vez que os relacionamentos conceituais básicos tenham sido definidos para um domínio, a extensão de um thesaurus multilíngüe ISO 5964 para outros idiomas é muito prático [Getty, 1997].

Um método complementar para a criação manual de base de conhecimento é o método da geração automática de thesaurus. Muitos são os trabalhos desenvolvidos: [Yuan, 1997] construiu um thesaurus no domínio da Tecnologia da Informação (IT) usando os recursos disponíveis o máximo possível, como thesauri existentes no mesmo domínio ou em domínios diferentes, além de dicionários. Dessas fontes de conhecimento, gerou-se uma lista de termos que mais tarde foi utilizada para identificar termos sobre IT. Depois usou-se um programa, o *MultiTes* [Multites, 1997], para organizar os relacionamentos e uni-lo a outros thesaurus. O MultiTes, software de construção e gerenciamento de thesaurus, facilita a criação, pesquisa, impressão e mantém diferentes tipos de vocabulários e thesauri em microcomputadores ou em rede local. MultiTes permite a manipulação de listas de termos, descritores, thesauri multilíngües, etc. O thesaurus IT tem um total de 8.456 termos; entre eles, 6.907 são únicos e 1.549 são sinônimos.

[Chen & Lin, 1996] desenvolveram um método algorítmico baseado em conceito para a classificação e recuperação da informação. A pesquisa utiliza um banco de dados bibliográfico multilíngüe que contém documentos técnicos principalmente em chinês, com alguns termos ocasionais em inglês. É gerado um espaço conceitual a partir da extração automática dos conceitos dos textos do banco

de dados e em seguida conceitos similares são agrupados pela análise da co-ocorrência dos conceitos nos textos.

Para o Worm Community System (WCS), uma implementação recente da tecnologia de sistemas de comunidade eletrônica, [Chen, 1995] propôs um método para gerar um thesaurus de domínio específico automaticamente através da análise de documentos armazenados usando listas de termos controlados adquiridos externamente (palavras-chaves), técnicas de indexação automática, e algoritmos de análise de agrupamento baseado em estatística. O thesaurus resultante capturou os conceitos de domínio específico e seus pesos, relacionamentos relevantes, e permitiu a atualização periódica, automática, de seus vocabulários e relacionamentos. Com uma interface simples, os usuários podem acessar o *WCS* usando seus próprios vocabulários e consultar o thesaurus, rico em semântica, para outros conceitos similares.

Em relação à recuperação de informação multilíngüe os primeiros experimentos sobre a sua viabilidade foram desenvolvidos por Salton [Salton, 1972]. As técnicas de análise e recuperação automática de textos incorporados ao sistema de recuperação textual SMART[1] foram usadas para processar documentos e consultas em inglês e alemão, utilizando um thesaurus multilíngüe (construído através de tradução). Salton concluiu que apesar da eficiência da recuperação variar entre coleções de documentos, "o processamento entre idiomas ... é quase tão eficaz quanto o processamento num único idioma".

O método mais tradicional para recuperação de informação multilíngüe, bastante utilizado em bibliotecas, usa um vocabulário controlado para

[1] SMART é um sistema de recuperação textual que utiliza o modelo de espaço vetorial.

indexar e recuperar, que consiste em achar para cada termo do thesaurus a sua tradução nos idiomas considerados. O TRANSLIB é um sistema de recuperação para um catálogo de biblioteca online que agrupa três idiomas: grego, espanhol e inglês [Oard, 1997b].

Um outro método para recuperação de informação multilíngüe é usar sistemas de tradução por máquina para traduzir as consultas e/ou a coleção de documentos. O SYSTRAN é um exemplo [Oard, 1997b].

O projeto de recuperação de informação multilíngüe européia (EMIR – European Multilingual Information Retrieval), guiado por Fluhr [Fluhr, 1995] do INSTN (Institut National des Sciences et Techniques Nucléaires) usou uma técnica de expansão de consulta. O objetivo do EMIR era estender o sistema de recuperação textual SPIRIT para múltiplos idiomas. O par de idiomas inicial foi inglês/francês, e foi mais tarde estendido para o alemão. O EMIR usa dicionários unilíngües e bilíngües habilitando o processamento de bancos de dados em qualquer domínio.

Observa-se nos trabalhos supracitados a tendência do uso de indexação automática para a geração de thesaurus, em conjunto ou não com dicionários bilíngües e sistemas de tradução por máquina.

No entanto, a construção automática não elimina, e até aumenta, os problemas de ambigüidade das palavras, ou seja, qual a interpretação correta a ser dada a uma determinada palavra em um certo contexto. Além disso, por mais poderoso que seja o software, as decisões humanas não podem ser substituídas, a fim de que o thesaurus funcione efetivamente.

Outrossim, o grande problema a ser resolvido em termos de recuperação refere-se à seguinte paráfrase: "como recuperar documentos que contêm expressões que não casam exatamente com aquelas estabelecidas na consulta? " [Fluhr, 1995].

## 1.3 OBJETIVO

Este trabalho tem como objetivo desenvolver um método para construção semi-automática de um thesaurus retangular multilíngüe, a partir de um conjunto de documentos multilíngües e dicionários unilíngües, partindo do pressuposto de que um documento corretamente indexado também será corretamente recuperado. Os idiomas considerados nestes trabalho são português, inglês, italiano e francês.

Outro objetivo foi transformar o protótipo de Construção Automática de Thesaurus, desenvolvido por [Ferneda, 1997], para atender ao método proposto. Desta transformação faz parte uma interface de pesquisa no thesaurus gerado a fim de possibilitar a recuperação de documentos.

## 1.4 ORGANIZAÇÃO DA DISSERTAÇÃO

Neste capítulo de *Introdução* apresentamos os aspectos que motivaram o desenvolvimento deste trabalho, além dos objetivos e trabalhos relacionados.

O capítulo 2, *Fundamentação Teórica*, apresenta os trabalhos que serviram de base para o desenvolvimento do método proposto nesta dissertação para a construção semi-automática de um thesaurus retangular multilíngüe.

O capítulo 3, *Método para Construção Semi-automática de um Thesaurus Retangular Multilíngüe*, descreve as fases envolvidas na construção do thesaurus, detalhando a extração de termos, a estrutura dos dicionários, o processamento das palavras compostas, a análise da co-ocorrência de termos até a geração de um thesaurus abstrato, independente de idioma. Outrossim, mostra que a base de documentos é gerada utilizando-se também o método de decomposição retangular de uma relação binária.

O capítulo 4, Protótipo para *Construção Semi-automática de um Thesaurus Retangular Multilíngüe - SISMULT*, apresenta os módulos de indexação e de consulta do protótipo.

No capítulo 5 são encontradas as considerações finais deste trabalho bem como propostas de trabalho relacionadas ao aperfeiçoamento da ferramenta desenvolvida.

# 2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica dessa pesquisa baseia-se nos trabalhos sobre extração de termos relevantes, métodos e algoritmos para a construção de thesaurus, e thesaurus multilíngue que são apresentados a seguir.

Termos relevantes (expressam o conteúdo do documento) são obtidos através do processo de indexação de documentos. O processo de indexação de documentos, manual ou automático, consiste em varrer o documento e identificar as palavras relevantes que possam relacionar o documento a um determinado assunto.

## 2.1 MÉTODO PARA EXTRAÇÃO DE TERMOS

Esta seção apresenta uma descrição de um método para extração de termos a partir de textos em linguagem natural, baseado nos trabalhos de Marie-Françoise Bruandet [Bruandet 1980a, 1980b, 1981, 1982, 1982a, 1985, 1989a, 1989b] e Attar [Attar,1977].

O método baseia-se na pesquisa de associações entre pares de palavras. Para cada par de palavras é calculado um valor que expressa a força de ligação entre elas, considerando-se para tanto:

- A proximidade contextual (distância) entre cada par de palavras que aparecem numa mesma frase;
- A categoria gramatical de cada uma das palavras;

- A freqüência com que o par de palavras aparece.

Durante a extração das palavras significativas do texto as palavras são reduzidas à sua forma normalizada (canônica). Por exemplo, para substantivos e adjetivos, a forma canônica considerada é o masculino-singular; para verbos, a forma infinitiva. A forma canônica de uma palavra é chamada *termo*.

Através de funções estatísticas calcula-se, para cada par de termos (x, y), uma medida que expressa a força de ligação entre x e y, através da seguinte fórmula:

$$M(x,y) = \frac{\sum_i \sum_j \frac{1}{d(wi(x), wj(y))}}{f(x,y)} \cdot \frac{\left(f(x,y)-1\right)^n}{f(x,y)^n}$$

Onde:
- f(x,y) é a frequência do par (x,y) no conjunto de documentos;
- $w_i(x)$ é a i-ésima ocorrência de um termo x no conjunto de termos extraídos dos documentos;
- $d(w_i(x), w_j(y))$ é a distância entre a i-ésima ocorrência de x e a j-ésima ocorrência de y.

Os valores dessas medidas são armazenados em uma matriz termo-termo. Em seguida constrói-se, a partir desta matriz, uma matriz binária termo-termo, processo em que as ligações mais fracas são eliminadas, de acordo com um limite mínimo estabelecido. A partir do grafo representado por essa matriz são extraídos os *cliques* (subgrafos completos máximos)[2], que representam as idéias contidas nos textos. A análise dos cliques mostra que as informações representativas são essencialmente veiculadas por substantivos e por certos adjetivos. Diversos

[2] Subgrafos cujos nós estão todos conectados entre si.

algoritmos para a extração de cliques de um grafo estão disponíveis no domínio da teoria dos grafos.

Em [Bruandet, 1989b] são apresentados alguns resultados obtidos em uma experimentação realizada com um texto contendo 15 capítulos, com um total de 70.350 palavras. Os resultados obtidos foram validados através da comparação com a indexação manual efetuada por documentalistas. Foram encontrados pela indexação automática 80% dos termos selecionados manualmente.

## 2.2 MÉTODO DA DECOMPOSIÇÃO RETANGULAR DE UMA RELAÇÃO BINÁRIA

Os métodos mais comuns para a construção de thesauri são os métodos estatísticos e lingüísticos. O método estatístico consiste em agrupar termos indexados similares e em seguida definir ligações entre eles. O problema deste método é a ausência de fundamentação matemática nas fórmulas usadas para o cálculo das similaridades entre os termos [Salton, 1989]. O problema do método lingüístico é a ambigüidade, o que torna necessária uma ferramenta para armazenar e gerenciar os termos lingüísticos, além de um especialista com conhecimento em lingüística.

[Ferneda, 1997] desenvolveu um construtor automático de thesaurus retangular utilizando um método - ***Método da Decomposição Retangular de uma Relação Binária*** [Gammoudi, 1993] - que tem boa fundamentação matemática, e tem sido útil em vários domínios, como Inteligência Artificial, Engenharia de

Software e Banco de Dados Documental. Com este método, a partir de uma matriz binária contendo a relação entre os termos extraídos dos documentos ou a relação entre termos e documentos, obtém-se um conjunto de retângulos ótimos que são os nós do thesaurus retangular.

A extração de retângulos de uma relação binária finita tem sido estudada por matemáticos no contexto da teoria dos reticulados, e tem provado ser de grande relevância para diversos campos da ciência da computação. Os retângulos ótimos extraídos de uma relação binária através da Decomposição Retangular são organizados sob forma de um grafo hierárquico através de uma relação de ordem parcial.

Uma decomposição retangular de uma relação binária finita R é um conjunto de retângulos ótimos que constitui uma cobertura mínima de R. Um retângulo ótimo é uma particularidade do retângulo máximo.

O método da Decomposição Retangular de uma Relação Binária consiste na decomposição da relação binária em *n* relações elementares e para cada relação elementar seleciona-se o conjunto de retângulos ótimos. Deste conjunto, escolhe-se o mínimo de retângulos ótimos que forme sua cobertura eliminando-se o máximo de elementos redundantes nos retângulos ótimos.

[Gammoudi, 1993] mostra que quando a cardinalidade de R é grande obtém-se um importante ganho em espaço de armazenamento, o que certifica que retângulos ótimos obtidos da decomposição de uma relação binária podem ser usados para representar grandes bancos de dados documentais.

### 2.2.1 Conceitos Básicos

*Definição 2.1: Relação Binária*

*Chama-se relação binária de um conjunto E em um conjunto F, ou simplesmente relação de E em F, todo subconjunto R do produto cartesiano E x F.*

Um elemento de uma relação binária R é denotado por (x,y). Indica-se por xRy o fato de um elemento x de E estar ligado a um elemento y de F através da relação R.

Uma *relação binária identidade*, designada por I, é uma relação tal que, se S é um conjunto qualquer, então $I(S)= \{ (x,x) \mid x \in S\}$.

Conjuntos associados a uma relação $R \subseteq ExF$:

I. $x.R= \{y \mid xRy\}$ Conjunto imagem de x

II. $R.y= \{ x \mid xRy\}$ Conjunto dos antecedentes de y

III. $dom( R ) = \{ x \mid \exists y: xRy\}$ Domínio de R

IV. $cod( R ) = \{y \mid \exists x: xRy\}$ Codomínio (imagem) de R

Sejam R e R' duas relações binárias de E em F. As seguintes operações são possíveis:

I. Interseção: $R \cap R'= \{(x,y) \in E \times F \mid xRy \ \& \ xR'y\}$, onde & é o símbolo de multiplicação lógica.

II. União: $R \cup R'= \{(x,y) \in E \times F \mid xRy \vee xR'y\}$, onde $\vee$ é o símbolo de adição lógica.

III. Inverso: $R^{-1} = \{(y,x) \in F \times E \mid xRy\}$

IV. Composição: Seja F=E, $R \circ R' = \{(x,y) \in E \times F \mid \exists t: xRt \ \& \ tR'y\}$

Propriedades de uma relação binária:

- Sejam R e R' duas relações binárias sobre um conjunto E. Diz-se que R é *mais determinista* que R' se e somente se $R^{-1} \circ R \subseteq (R')^{-1} \circ R'$. Isto significa que quanto mais imagens uma relação associar a uma entrada, *menos determinista* ela será;

- Para duas relações binárias R e R' temos que, $(R \circ R')^{-1} = (R')^{-1} \circ R^{-1}$

Na teoria dos grafos uma relação binária R de E em F define os arcos de um grafo bipartido sobre E x F.

***Definição 2.2: Retângulo de uma Relação Binária***

---

*Seja R uma relação binária definida de E em F. Um retângulo de R é um par de conjuntos (A,B) tal que $A \subseteq E$, $B \subseteq F$ e $A \times B \subseteq R$. A é o domínio do retângulo enquanto B é o seu codomínio.*

---

O *fechamento retangular* de uma relação R é a relação $R^{++} = dom(R) \times cod(R)$.

Pode-se dizer que um retângulo é um subgrafo bipartido completo (ou clique) do grafo $(E \cup F, R)$. Um subgrafo é bipartido se seus vértices podem ser particionados em dois subconjuntos $V_1$ e $V_2$ de tal modo que dois vértices do mesmo

conjunto não sejam adjacentes. Um subgrafo é completo se todos os seus pares de vértices são adjacentes [Furtado, 1973]. Então dada a relação $R \subseteq E \times F$ e dois conjuntos A e B, A={a,b,c} e B={e,f}, dizer que (A,B) é um retângulo de R, significa que nenhum elemento de A (B) está relacionado a outro elemento de A (B) e que todo elemento de A(B) está relacionado a um elemento de B(A).

***Definição 2.3: Retângulo Máximo***

---

*Seja R uma relação binária definida de E em F. Um retângulo (A,B) de R é dito máximo se e somente se, para todo retângulo (A',B')*

$$A \times B \subseteq A' \times B' \subseteq R \rightarrow A = A' \text{ e } B = B'.$$

---

***Proposição 2.1:***

---

*Seja R uma relação binária finita e (x,y) $\in$ R. A união dos retângulos de R que contêm o elemento (x,y) é igual à relação elementar*

$$\phi_R(x,y) = I(y.R^{-1}) \circ R \circ I(x.R).$$

---

A Figura 2.1(d) é um exemplo de uma relação elementar contendo o elemento (a,1) da relação inicial R, ilustrada pela Figura 2.1(a).

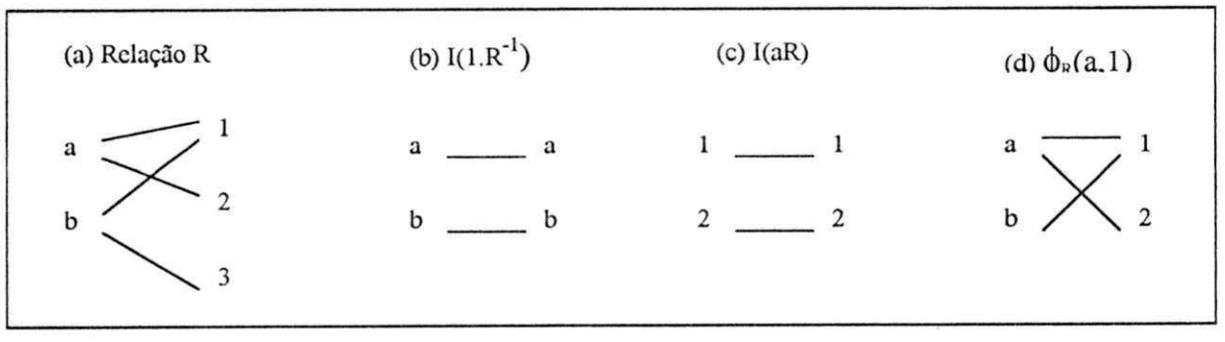


*Figura 2.1: Ilustração da noção de Relação Elementar*

***Definição 2.4: Ganho em espaço de armazenamento***

---

*O ganho em espaço de armazenamento de um retângulo qualquer RE= (A,B) é medido da seguinte forma:*

$$g(RE) = [Card(A) \times Card(B)] - [Card(A) + Card(B)]$$

---

onde Card= cardinalidade do conjunto.

Desde que g(RE) ≥ 0, o que é atingido quando Card(A) > 1 e Card(B) > 1, obtemos um ganho apreciável em espaço de armazenamento da informação. Se Card(A) >2 e Card(B) > 2 então g(RE) > 0 e cresce em função de Card(A) e Card(B). No entanto, o desperdício (ganho negativo) não pode jamais ultrapassar o valor 1, atingido para Card(A) =1 ou Card(B) =1. Pode-se representar todos os elementos do retângulo RE pelos pares (a,n) e (n,y), onde **n** é uma constante que identifica o retângulo RE (n=1 na Figura 2.2b), **a** representa um elemento qualquer do conjunto A, e y representa um elemento qualquer do conjunto B. A Figura 2.2, ilustra os dois modos de representação possíveis para um mesmo retângulo RE= ({a,b,c}, {x,y,z}), mostrando que podemos representar os 9 pares iniciais da relação através de 6 pares, obtendo assim uma economia no armazenamento da informação de 3 pares.

***Definição 2.5: Retângulo ótimo***

---

*Um retângulo máximo contendo um elemento (x,y) de uma relação R é dito ótimo se ele produz o máximo de ganho entre todos os retângulos máximos que contêm (x,y).*

---

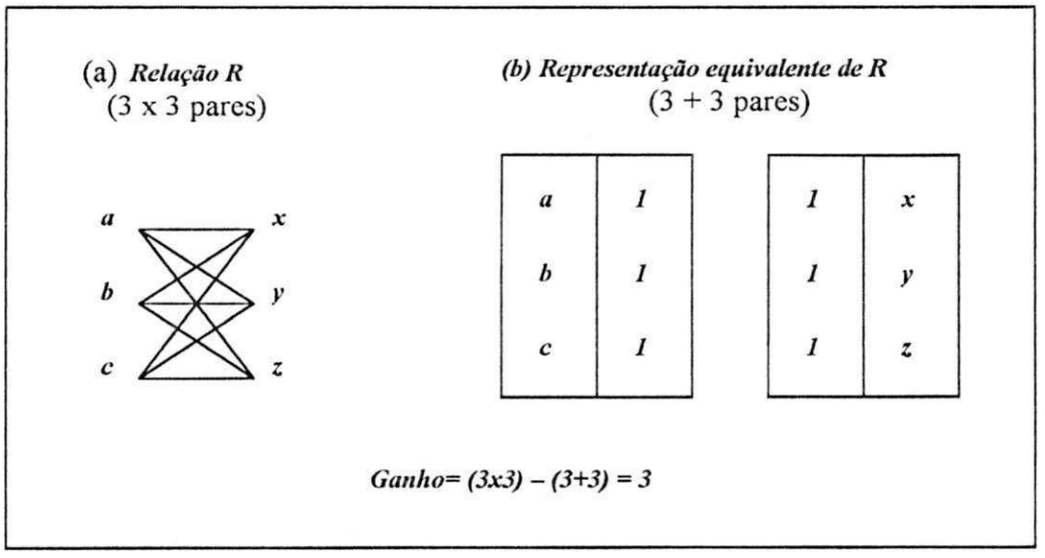


*Figura 2.2: Representações equivalentes de um mesmo retângulo*

A Figura 2.3(a) apresenta um exemplo de uma relação R, e as Figuras 2.3(b), 2.3(c) e 2.3(d) representam três retângulos máximos que contêm o elemento (y,3), cujos ganhos em espaço de armazenamento são respectivamente 1, 0 e -1. Logo, o retângulo ótimo que contém o elemento (y,3) de R é o retângulo ilustrado pela Figura 2.3(b).

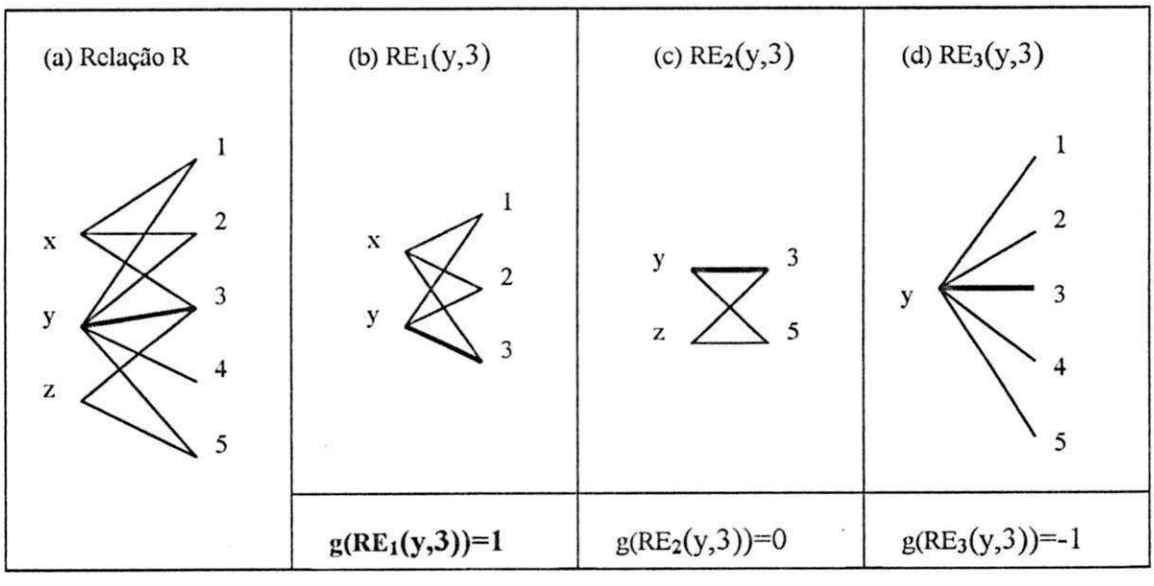


*Figura 2.3: Ilustração de Retângulo Ótimo*

***Definição 2.6: Cobertura de uma relação***

---

*Chama-se cobertura de uma relação R um conjunto de retângulos C= {$RE_1$, $RE_2$, ..., $RE_n$} de R tal que todo elemento (x,y) de R pertence a pelo menos um dos retângulos de C.*

---

***Definição 2.7: Cobertura mínima de uma relação***

---

*Uma cobertura C= {$RE_1$, $RE_2$, ..., $RE_n$} de uma relação R é dita mínima se nenhum subconjunto próprio de C é uma cobertura.*

---

***Definição 2.8: Relação retangular***

---

*Seja R uma relação binária definida sobre um conjunto E e (A,B) um retângulo. A relação A x B $\subseteq$ R é chamada relação retangular associada ao retângulo (A,B) de R. A é o domínio da relação retangular e B é seu codomínio.*

---

Existe uma relação biunívoca entre os retângulos ($A_i$,$B_i$) e as relações retangulares $A_i$ x $B_i$, excetuando-se o caso em que $A_i$= $\varnothing$ ou $B_i$= $\varnothing$, que corresponderiam à relação retangular $\varnothing$. Para facilitar a leitura do texto utilizar-se-á a nomenclatura "retângulo (A,B) contendo um elemento (x,y) de uma relação R" no lugar de "relação retangular A x B de R, associada ao retângulo (A,B), e contendo um elemento (x,y) de R" que seria o mais preciso.

***Definição 2.9: Grafo Retangular***

---

*Seja a relação "$\leq\leq$" definida sobre o conjunto de retângulos de uma relação binária R, como segue:*

$$\forall (A_1,B_1) \text{ e } (A_2, B_2) \text{ dois retângulos de } R:$$

$$(A_1,B_1) \leq\leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \text{ e } B_2, \subseteq B_1.$$

*Chamamos (R, $\leq\leq$) de Grafo Retangular.*

---

"$\leq\leq$" é uma relação de ordem parcial pois ela é:

- Reflexiva: $x \leq\leq x$
- Anti-simétrica: $x \leq\leq y$ e $y \leq\leq x \Rightarrow x = y$
- Transitiva: $x \leq\leq y$ e $y \leq\leq z \Rightarrow x \leq\leq z$

***Definição 2.10: Reticulado.***

---

*Seja (R,<) um conjunto R com uma relação de ordem <. Diz-se que (R, <) é um reticulado se e somente se todo subconjunto $X \subseteq R$ admite um menor limite superior e um maior limite inferior.*

---

***Proposição 2.2:***

---

*Seja R uma relação binária definida sobre um conjunto E e G o conjunto dos retângulos ótimos de R ordenados pela relação "$\leq\leq$". (G, $\leq\leq$) é um reticulado com um nódulo ínfimo (o,E) e um nódulo supremo (E,o).*

---

### 2.2.2 Relações semânticas

É possível distinguir três tipos de relações num thesaurus retangular: relações hierárquicas (termo genérico e termo específico), relações de equivalência (termos sinônimos e pseudo-sinônimos) e relações não-hierárquicas (termos vizinhos) [Aitchison, 1979].

Estas relações semânticas são utilizadas na recuperação de informação para identificar no grafo retangular quais os retângulos que contêm informação mais genérica, ou mais específica, ou mesmo relacionada, ao retângulo que contém o termo de pesquisa.

As relações hierárquicas são definidas introduzindo-se a noção de generalização e especificação entre os retângulos, no sentido vertical.

***Definição 2.11: Retângulo genérico e específico***

*Seja $RE_i = (A_i, B_i)$ e $RE_j = (A_j, B_j)$ dois retângulos ótimos de R. $RE_i$ é genérico em relação a $RE_j$ ($RE_j$ é específico em relação a $RE_i$) se:*

$$(A_i, B_i) \leq\leq (A_j, B_j) \Leftrightarrow (A_i \subseteq A_j) \text{ e } (B_j \subseteq B_i).$$

Ou seja, se Card($A_i$) é menor do que a Card($A_i$), então $A_i$ é mais genérico do que $A_j$.

*Definição 2.12: Grau de generalidade/especificidade*

---

*Seja $RE_i = (A_i, B_i)$ e $RE_j = (A_j, B_j)$ dois retângulos ótimos de R, tal que existe uma ligação hierárquica entre $RE_i$ e $RE_j$, ou seja, $Card(A_i) \neq Card(A_j)$. O grau de generalidade/especificidade é definido por:*

$$G_{g,e}: R_{Rótimo} \times R_{Rótimo} \rightarrow [0,1]$$

$$G_{g,e}(RE_i, RE_j) = \frac{1}{ABS(Card(A_{ij}) - Card(Aj))}$$

---

Por exemplo, na Figura 2.4 o grau de genericidade entre os retângulos $R_8$ e $R_2$ é dado por $G(R_8,R_2) = 1/ABS(Card(R_8.termo)-Card(R_2.termo))= 1/(3-1)=0,5$, e $R_8$ ($R_2$) é mais específico(genérico) do que $R_2$ ($R_8$).

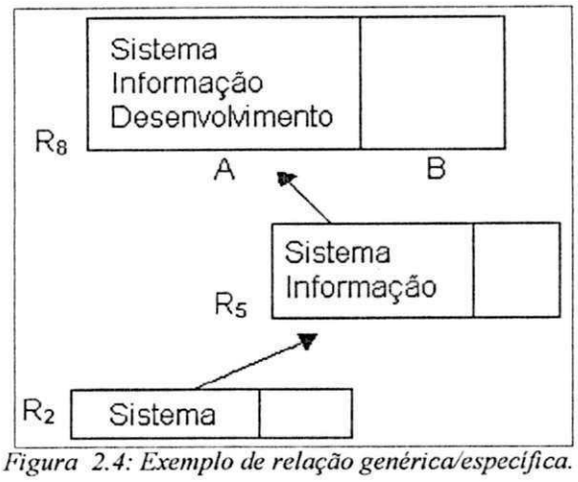


*Figura 2.4: Exemplo de relação genérica/específica.*

As relações de equivalência incluem os sinônimos e pseudo-sinônimos. A relação sinônimo é uma relação de equivalência entre os termos relevantes extraídos dos documentos. Cada termo pertencente ao domínio ou codomínio representa uma classe de equivalência, que possui um mesmo conceito.

Pseudo-sinônimos referem-se aos termos do domínio ou codomínio de um determinado retângulo.

Na Figura 2.5 temos que os termos *Informação, Recuperação* e *Documento* são pseudo-sinônimos e representantes dos seus respectivos sinônimos: *Fatos e dados, busca e consulta, e relatório, formulário e artigo.*

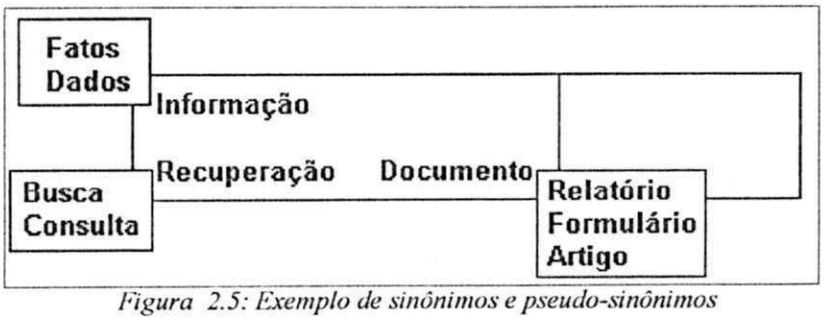


*Figura 2.5: Exemplo de sinônimos e pseudo-sinônimos*

As relações não-hierárquicas expressam associações entre retângulos, permitindo assim, indicar certas analogias ou aproximações entre dois retângulos que se sobrepõem parcialmente [Ferneda, 1997].

***Definição 2.13: Retângulos vizinhos***

---

*Seja* $RE_i = (A_i, B_i)$ *e* $RE_j = (A_j, B_j)$ *dois retângulos ótimos de R.* $RE_i$ *é vizinho a* $RE_j$ *se e somente se as seguintes condições forem verificadas:*

$$A_i \cap A_j \neq 0 \text{ ou } B_i \cap B_j \neq 0$$

$$(A_i, B_i) \not\leq (A_j, B_j) \text{ ou } (A_j, B_j) \not\leq (A_i, B_i)$$

---

***Definição 2.14: Grau de Vizinhança***

---

*Vizinhança:* $R_{ótimo} \times R_{ótimo} \rightarrow [0,1]$

$\forall (RE_i, RE_j) \in R_{ótimo} \times R_{ótimo}$ *tal que* $RE_i$ *é vizinho de* $RE_j$.

$$Vizinhança(RE_i, RE_j) = \frac{Card[Dom(RE_i) \cap Dom(RE_j)]}{Card[Dom(RE_i) \cup Dom(RE_j)]}$$

---

Dois retângulos são ditos distintos ou não-vizinhos se o grau de vizinhança entre eles é zero. Inversamente, quanto mais o grau de vizinhança de dois retângulos se aproxima de 1, mais esses dois retângulos são vizinhos permitindo recuperar documentos que possuem grande similaridade.

A Figura 2.6 mostra que os dois retângulos são vizinhos porque a interseção entre eles não é vazia e porque não estão relacionados hierarquicamente.

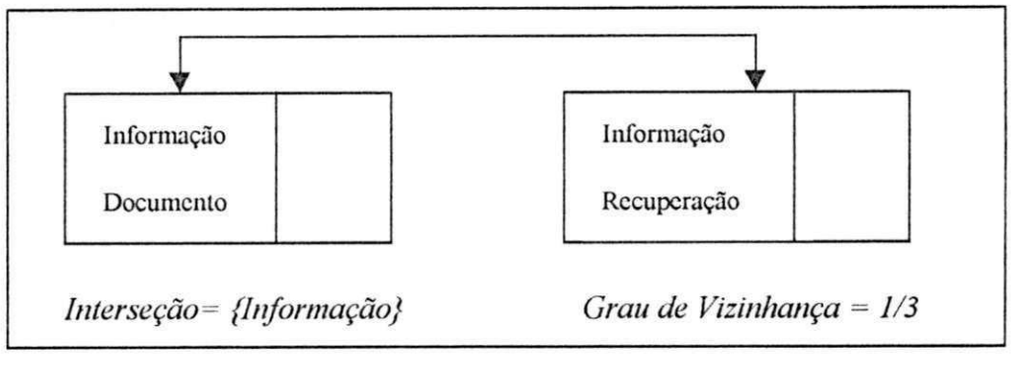


*Figura 2.6: Relação de vizinhança*

## 2.3 ALGORITMO DE PINTO

O Algoritmo de GODIN [Godin, 1986], um dos mais conhecidos para a construção de grafo retangular, gera o grafo retangular de maneira incremental a

partir de uma relação binária, sendo o reticulado organizado em níveis hierárquicos em função da cardinalidade dos domínios dos retângulos [Pinto, 1997]. Cada linha da matriz, representação da relação binária, é processada da seguinte maneira:

1. Cada linha da matriz é transformada em um novo retângulo máximo;
2. Obtém-se a cardinalidade do novo retângulo;
3. Insere-se o novo retângulo no grafo obedecendo o nível de cardinalidade;
4. Verifica-se a existência de relação de ordem parcial entre o novo retângulo e os existentes, em todos os níveis, para efetuar as ligações.

O resultado do estudo da complexidade do Algoritmo de GODIN é da ordem $O(n^2)$ para inserir um novo retângulo [Pinto, 1997].

O *Algoritmo de PINTO* [Pinto, 1997], reúne as características relevantes do método da Decomposição Retangular de uma Relação Binária e do Algoritmo de GODIN, isto é, reduz o número de retângulos redundantes através na noção de ganho em espaço de armazenamento e é incremental.

O Algoritmo de PINTO obtém uma cobertura reduzida da relação binária, buscando assim eliminar retângulos redundantes, da seguinte forma: quando um retângulo originado da interseção entre um novo retângulo a ser inserido e um retângulo pertencente ao grafo é gerado, calcula-se o seu ganho em armazenamento verificando se ele será ou não inserido no grafo. Este algoritmo apresenta complexidade $O(n^{3/2})$ para inserir um novo retângulo no grafo [Pinto, 1997]. A redução do número de retângulos é da ordem de 15%.

Algoritmo de PINTO:

1. Obter a cardinalidade C do domínio do novo retângulo RE (que é uma relação binária T x D, onde T é o conjunto de termos e D o conjunto de documentos) representando o novo documento a ser inserido;
2. Verificar a existência de cardinalidade igual a C na lista de cardinalidades (LC);
3. Inserir o novo retângulo na cardinalidade correspondente;
4. Para todo retângulo $RE_i$ do grafo, com cardinalidade diferente de C, verifica-se a existência de relação de ordem parcial, caso contrário verifica-se a existência de interseção do domínio de $RE_i$ com o domínio de RE procedendo a inserção do retângulo derivado da mesma, isto é, de um novo retângulo que contém a união dos codomínios de RE e $RE_i$ e a interseção dos domínios de RE e $RE_i$.

## 2.4 THESAURUS MULTILÍNGÜE

[Sosoaga, 1991] define um thesaurus como sendo um sistema de classificação onde a estrutura interna é composta por um conjunto de conceitos relacionados entre si pelos relacionamentos usuais da biblioteconomia (termo relacionado, termo genérico, termo específico), e a estrutura externa como sendo um conjunto de palavras correspondentes aos conceitos. Ou seja, um thesaurus pode ser um sistema definido por

$$\mathbf{Th= ( V, n, r; L, t)}$$

onde V é o conjunto de conceitos; n e r são duas relações diferentes definidas em V; L é um conjunto de palavras num certo idioma e t é uma aplicação que projeta L em V.

A função t associa a cada termo um conceito dependendo do contexto. A inversa da função t induz em L uma relação de equivalência, isto é, o conjunto de termos que possuem o mesmo conceito para um determinado contexto. Desta forma, podemos decompor t em duas funções, $t_0$ e $t_1$, onde $t(x,c) = t_1(t_0(x,c))$. A função $t_0$ é responsável pela escolha de um termo canônico da classe de equivalência e a função $t_1$ é uma função injetiva que determina o conceito abstrato. Ou seja,

$$LxC \xrightarrow{t} V \Leftrightarrow LxC \xrightarrow{t0} L_0xC \xrightarrow{t1} V$$

onde $L_0$ é o dicionário dos representantes das classes de equivalência. Esta decomposição corresponde ao processo de escolha dos conceitos que indexam um documento processado.

As relações n e r representam os relacionamentos usuais para cada termo: termo específico (NT), termo relacionado (RT). As suas relações inversas também estão incluídas: termo genérico (BT) inverso de NT, RT inverso de RT, que é simétrico. A relação n é uma relação de ordem parcial e r é uma relação de equivalência entre conceitos.

Desde que o conjunto de conceitos de um thesaurus (estrutura interna) é único e independente das palavras a ele ligadas, isto é, não depende nem dos documentos nem do idioma, é possível estender a definição de thesaurus unilíngüe

para thesaurus multilíngüe adicionando à sua estrutura unilíngüe um ou vários dicionários em diversos idiomas, bem como as respectivas funções que os projetam no conjunto conceitual, para um determinado contexto.

Logo, um thesaurus multilíngüe de ordem k é definido por:

$$ThM= (V, n, r; L_1, t_1, L_2, t_2, ..., L_k,, t_k)$$

onde:

V= Conjunto de conceitos

n, r = Relações definidas em V

$L_i$= Dicionário no idioma i, i=1,...,k

$t_i$= Função que projeta $L_i$ em V.

Um thesaurus multilíngüe, portanto, é um sistema de classificação composto por um conjunto de conceitos relacionados ao qual estão ligados conjuntos de termos expressos em diferentes idiomas. A Figura 2.7 apresenta uma representação para um thesaurus bilíngüe.

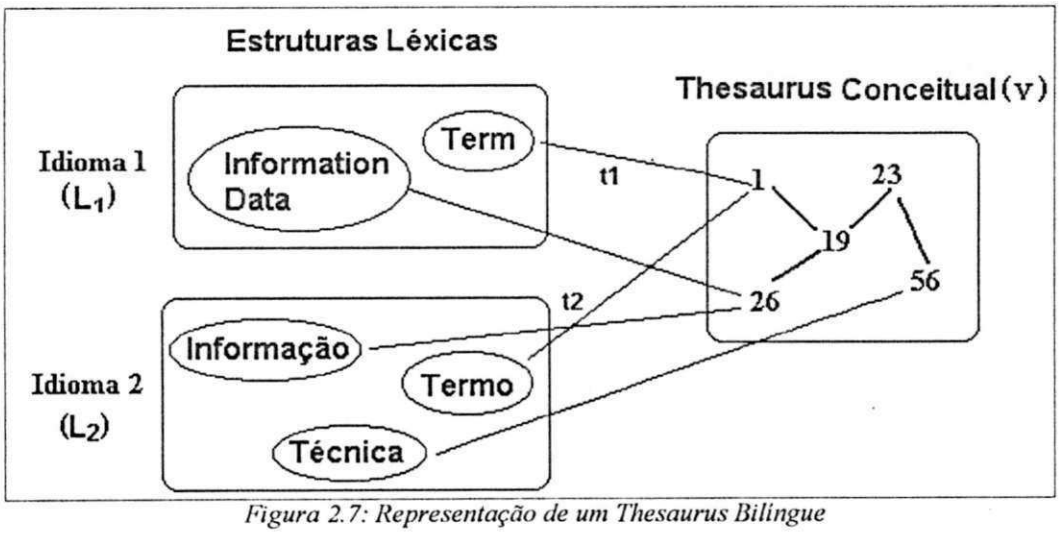


*Figura 2.7: Representação de um Thesaurus Bilíngue*

# 3 MÉTODO PARA CONSTRUÇÃO SEMI-AUTOMÁTICA DE THESAURUS RETANGULAR MULTILÍNGÜE

Este capítulo tem como objetivo apresentar um método para construção semi-automática de thesaurus retangular multilíngüe a partir de um conjunto de documentos em linguagem natural. Este método reúne os trabalhos apresentados no Capítulo 2, estendendo e/ou acrescentando alguns conceitos quando necessário.

O método consiste na extração de termos relevantes do conjunto de documentos (método de Bruandet modificado) e identificação de combinações que são importantes para representar os documentos. Para tais combinações, calcula-se a força de ligação que é armazenada numa matriz binária, a partir da qual são gerados os retângulos ótimos utilizando o método de decomposição retangular de uma relação binária [Gammoudi, 1993], que são inseridos no thesaurus retangular multilíngüe através do algoritmo incremental de Pinto [Pinto, 1997].

São utilizados dicionários unilíngües e a ligação entre os idiomas ocorre pelo termo abstrato ou conceito. O que chamamos de termo pode ser uma palavra simples ou composta.

O conceito de Thesaurus Retangular Multilíngüe é apresentado a seguir, assim como a descrição completa do método.

## 3.1 THESAURUS RETANGULAR MULTILÍNGÜE

Conforme Sosoaga [Sosoaga, 1991], Capítulo 2, thesaurus multilíngüe é um sistema de classificação composto por um conjunto de conceitos relacionados ao qual estão ligados conjuntos de termos expressos em diferentes idiomas. No entanto, como desejamos construir um thesaurus multilíngüe a partir de documentos em linguagem natural, a noção de *"contexto"* tornar-se importante para dirimir a ambigüidade das palavras no processo de indexação. Assim, a definição de thesaurus multilíngüe de [Sosoaga, 1991], de ordem k, foi estendida para incluir a idéia de contexto da seguinte forma:

$$ThM = (V, n, r; L_1, C_1, t_1, L_2, C_2, t_2, \ldots, L_k, C_k, t_k)$$

onde V= Conjunto de conceitos;

n, r = Relações definidas em V;

$L_i$= Dicionário no idioma i, i=1,...,k;

$C_i$= Conjunto de Contextos no idioma i,

$t_i$= Função que projeta $L_i$ x $C_i$ ,em V.

Nem sempre existe, para um conceito $t_i(l_i,c_i)$ de um idioma i, um termo $l_j$ correspondente em outro idioma j, tal que $t_j(l_j,c_j) = t_i(l_i,c_i)$, isto é, embora os conceitos em V devam ter equivalentes em outros idiomas, podem existir conceitos com equivalentes em uns idiomas e não em outros.

Utilizando a definição de thesaurus multilíngüe com contexto e a definição de thesaurus retangular apresentada no Capítulo 2, dizemos que um thesaurus retangular multilíngue com contexto é um thesaurus multilíngüe onde o conjunto de conceitos é representado por um conjunto de retângulos ótimos. Cada retângulo ótimo, como já visto, é composto por um *domínio* e um *codomínio*. O domínio está relacionado ao conjunto de conceitos, independente de idioma e dependente de contexto, que representam os termos de indexação dos documentos, e o codomínio é o conjunto dos documentos indexados pelos conceitos contidos no domínio. A Figura 3.1 apresenta graficamente um thesaurus retangular bilíngüe.

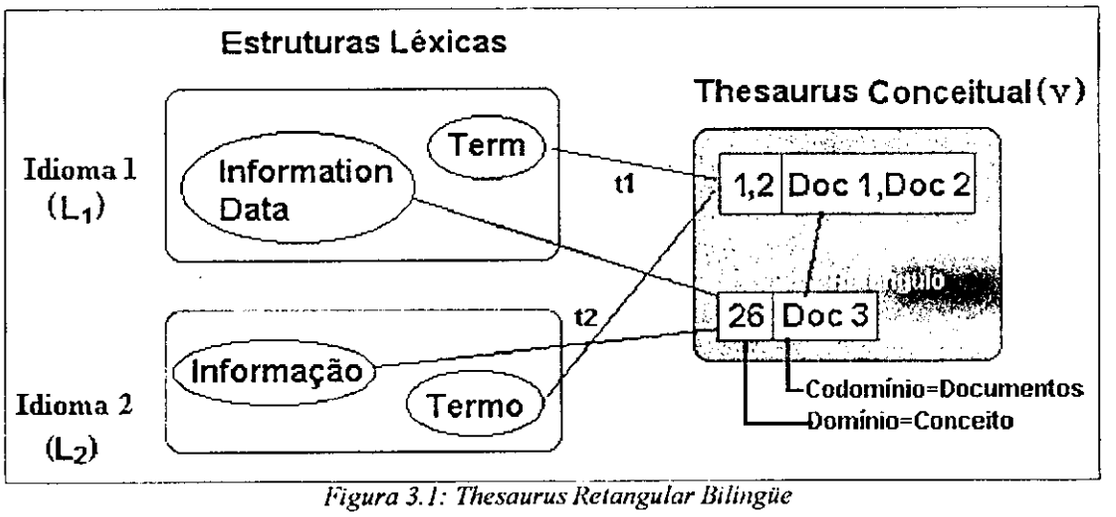


*Figura 3.1: Thesaurus Retangular Bilíngüe*

## 3.2 DESCRIÇÃO DO MÉTODO

O método semi-automático para construção de um thesaurus retangular multilíngüe envolve as seguintes etapas:

- Extração de termos de um ou mais documentos, determinando os conceitos abstratos, através da utilização de um dicionário unilíngüe e interação com o

usuário. Um termo extraído do documento, que não for uma stopword, é reduzido à forma canônica, e se tiver homônimos, o usuário decide, pelo contexto, qual conceito é adequado.

- Geração de retângulos ótimos a partir de uma matriz binária conceito-conceito (ou conceito-documento quando atualizando a base de documentos);
- Construção/atualização do thesaurus abstrato conceito-conceito (ou conceito-documento) existente pela incorporação dos retângulos ótimos usando o algoritmo incremental.

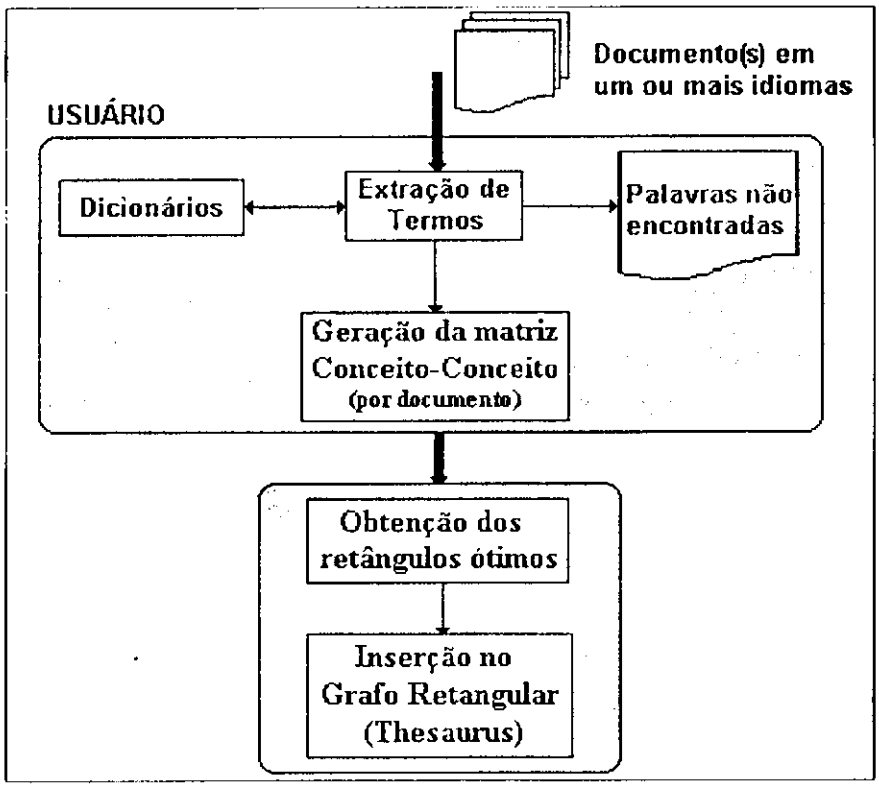


*Figura 3.2: Etapas para a Construção de um Thesaurus*

### 3.2.1 EXTRAÇÃO DE TERMOS

A construção de thesauri a partir de uma coleção de documentos eletrônicos em linguagem natural inicia com a extração de termos relevantes contidos nos documentos. A extração utiliza o método de indexação semi-automática que permite ao usuário experiente decidir, no caso de palavras ambíguas, qual conceito utilizar para determinado contexto. Em seguida, considera-se pares de palavras contidas numa mesma frase e calcula-se a força de ligação entre elas através do método de Bruandet (Capítulo 2).

O método de indexação semi-automática ocorre em duas fases: 1) fase de pré-indexação automática do texto a fim de selecionar termos relevantes; 2) fase de diálogo, durante a qual o usuário avalia os termos extraídos na fase anterior [Gammoudi, 1993]. Apesar de utilizar a idéia da indexação semi-automática, no tocante ao diálogo com o usuário, o processamento não é executado em duas fases e sim concomitantemente, pois o contexto em que o termo está inserido é importante e determina se o mesmo é ou não relevante (Figura 3.3).

Para identificar referências a informações relevantes num texto, um dicionário é usado como principal fonte de conhecimento. Como o objetivo deste trabalho é a construção de um thesaurus multilíngüe, o conjunto de documentos pode conter documentos em vários idiomas, o que implica na existência de vários dicionários unilíngües.

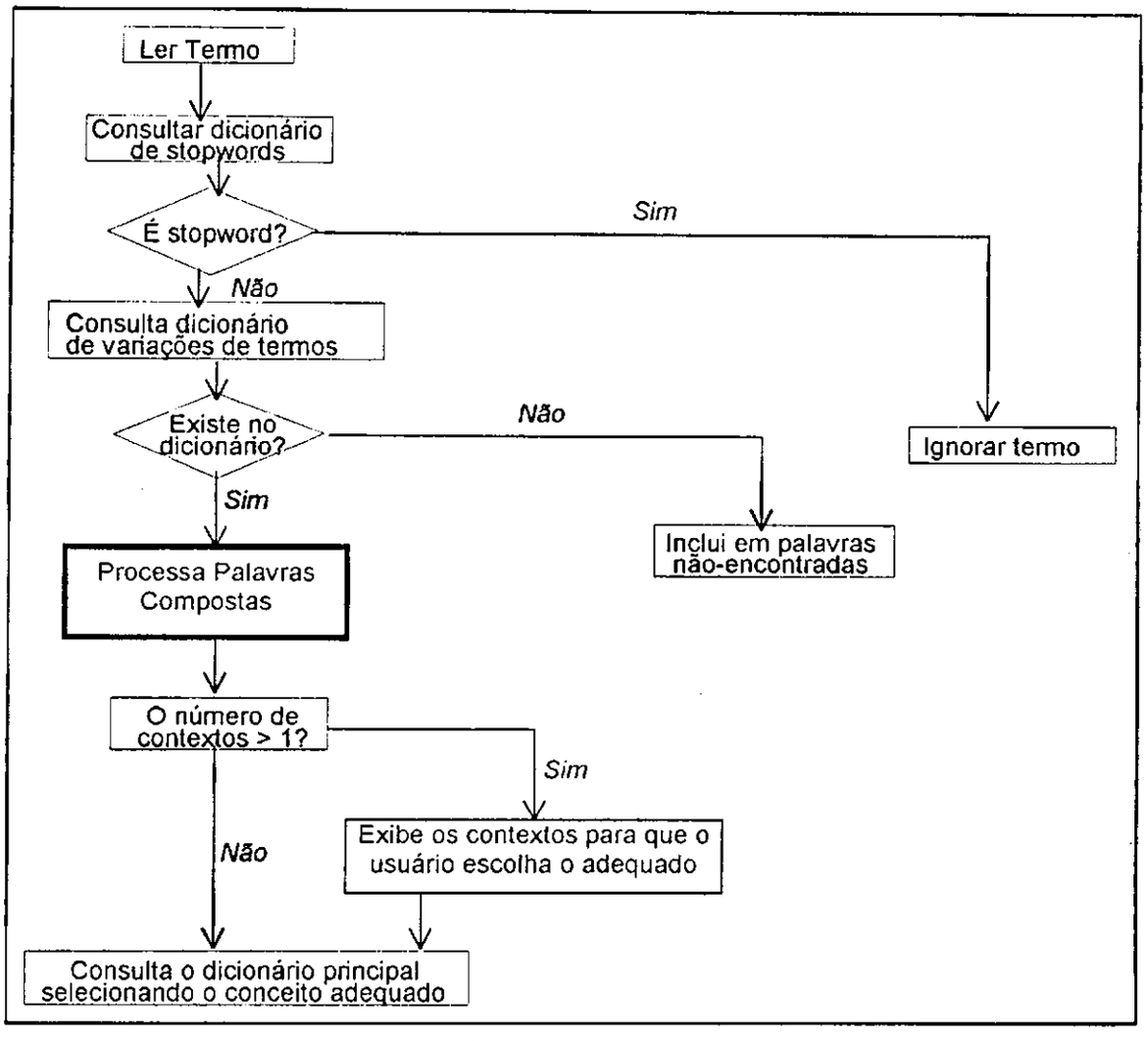


*Figura 3.3: Algoritmo simplificado do processo de indexação semi-automático.*

Após a identificação dos termos relevantes, utiliza-se uma técnica de agrupamento, por esta apresentar maior riqueza semântica que a estrutura de listas invertidas [Pinto, 1997], para identificar a associação entre pares de conceitos. A análise das co-ocorrências entre pares de palavras permite estabelecer índices estatísticos que representam a força de ligação entre seus pares e, a partir dos valores encontrados, mapear o estado de uma área do conhecimento num determinado momento. Das técnicas existentes escolheu-se a técnica de *cliques* que produzem classes que apresentam os relacionamentos mais fortes entre seus conceitos [Kowalski,1997].

A fundamentação para esta técnica de cliques pode ser encontrada nos trabalhos desenvolvidos por Marie-Françoise Bruandet do *Laboratoire Génie Informatique de Grenoble* [Bruandet 1980a, 1980b, 1981, 1982a, 1982b, 1985, 1989a, 1989b], que, por sua vez, se baseou em [Attar, 1977].

### 3.2.1.1 Processando o Texto

A primeira etapa da extração de termos é o processamento dos textos para selecionar as palavras que pertencem às categorias gramaticais que veiculam informações representativas: *substantivo*, *adjetivo* e *verbo*. As palavras passam por um processo de normalização onde os substantivos e adjetivos são reduzidos à sua representação no masculino-singular, e os verbos à sua forma no infinitivo através de consulta aos dicionários.

Analisa-se o texto por frases pelo fato destas serem a menor unidade representando uma idéia. Para tanto é necessário detectar o fim de frase, observando-se casos especiais como as elipses (...) e as aspas ("), isto em relação ao ponto. Outros caracteres são considerados também como delimitadores de fim de frase: sinal de exclamação, sinal de interrogação, final de parágrafo [Kavanagh, 1995]. Para cada palavra encontrada registra-se a sua localização, isto é, o documento que a contém, a frase, e a posição em que se encontra na frase.

A seleção de quais palavras serão utilizadas na construção do thesaurus é feita através de pesquisa ao dicionário do idioma do texto sendo processado. O dicionário é considerado a principal fonte de conhecimento devendo nele, portanto, ser encontrado o conceito para cada termo.

O resultado desta etapa é uma lista de conceitos sobre a qual realizam-se operações estatísticas e uma lista de palavras que não foram encontradas no dicionário e que não são *stopwords* (palavras sem valor de indexação). Esta lista de palavras é apresentada ao usuário para atualização dos dicionários.

### 3.2.1.2 Palavras Compostas

Em linguagem natural são freqüentes as situações em que certas seqüências de palavras têm um significado diferente daquele que seria inferível a partir dos significados das partes, surgindo portanto a necessidade de pré-agrupamento das palavras em **palavras compostas** para um tratamento correto destas seqüências. Além disso, uma palavra simples em um idioma pode ser composta em outro.

Palavras compostas são usadas para designar genericamente seqüências de palavras que têm características não dedutíveis das características das partes constituintes [Pinto e Almeida, 1995]. Exemplos: pezinhos de lã (idiomatismos) e máquina de escrever.

Não faz sentido que as palavras constituintes de uma palavra composta sejam agrupadas com a palavra isolada. Exemplo: a palavra "gato" no composto "gato pingado" não está relacionada com o mamífero gato, por isso quem pesquisar informação sobre gatos não tem interesse em encontrar a entrada relativa a "gato pingado". O significado do composto, por definição, não é dedutível a partir da semântica das palavras que o constituem, logo a única hipótese do tratamento

semântico correto de palavras compostas é ter a indexação explícita das mesmas [Pinto e Almeida, 1995]. Por outro lado o reconhecimento de uma palavra composta não se limita a uma simples comparação de cadeias de caracteres, já que as próprias palavras que compõem a palavra composta podem ser flexionadas.

Como a detecção e tratamento das palavras compostas baseia-se em dicionários e na recuperação da informação pode haver utilização de palavras compostas, surge a necessidade de existência de entradas das mesmas nos dicionários com conceitos próprios dependendo do contexto. Por exemplo, para encontrar documentos sobre "base de dados" interessa também "bases de dados", mas não interessa apenas "base" ou apenas "dados". A palavra composta "base de dados" tem um conceito diferente da palavra "base" e da palavra "dados".

O processamento das palavras compostas é semelhante ao processamento de palavras simples no tocante à identificação do conceito.

A noção de palavra composta não deve ser confundida com conceito composto que representa a composição de termos significativos em uma frase e é a base para as ligações hierárquicas na construção do thesaurus.

#### 3.2.1.3 O Dicionário

Para um thesaurus multilíngüe é necessário especificar os sinônimos entre os idiomas. Esta especificação, no entanto, não precisa ser completa, já que alguns termos não possuem traduções diretas em um outro idioma. Além disso, o que fazer com as palavras que possuem mais de um significado em um outro idioma?

As palavras sinônimas são agrupadas em classes, e para cada classe é atribuído um *conceito*. Este conceito é abstrato, isto é, independente do idioma, representando assim todos os sinônimos em todos os idiomas de um certo termo. Por outro lado, uma determinada palavra pode pertencer a mais de uma classe de sinônimos, e qual conceito usar depende do contexto desta palavra no texto analisado. Todos os contextos de uma determinada palavra constam no dicionário e o usuário deve decidir a qual contexto a palavra se refere.

A utilização de um conceito para representar os termos extraídos dos documentos com mesmo significado reforça a característica multilíngüe do thesaurus, pois o conceito independe dos documentos analisados, consequentemente, independe também do idioma (Capítulo 2). Isto também possibilita que, na consulta a documentos, o usuário utilize qualquer sinônimo de um certo conceito para obter todos os documentos relevantes.

Para cada idioma considerado tem-se um dicionário principal associado e uma lista de stopwords. Cada dicionário[3], cuja estrutura é apresentada a seguir, pode ser considerado um thesaurus, e foi construído manualmente para este trabalho.

(a) termo canônico

(b) A categoria Gramatical do termo – substantivo, adjetivo ou verbo

(c) Contexto do termo

(d) Conceito abstrato

(e) Identificação do representante

[3] Exemplos deste dicionário pode ser encontrado no Apêndice A.

(f) Apontador para uma lista de termos relacionados (termos relacionados conceitualmente mas não hierarquicamente. Um exemplo é a relação parte/todo).

As *stopwords* são palavras de função comum, de alta freqüência nos textos, que são insuficientes para representar conteúdo. Palavras de função gramatical como preposição, conjunção, artigo e pronome, exibem aproximadamente freqüências iguais de ocorrência em todos os documentos de uma coleção [Salton, 1989]. São palavras sem valor de indexação e portanto, sem valor de pesquisa, sendo utilizadas para otimizar a indexação eliminando a indexação de palavras que nunca serão pesquisadas. É possível também, através das stopwords, eliminar a indexação de qualquer palavra que o usuário julgue não ser importante para o seu propósito [Bussmann, 1995].

Além do dicionário principal, existe um dicionário auxiliar que contém as variações ortográficas dos termos, uma indicação se o termo pode ser composto e um apontador para o dicionário principal identificando a que classe conceitual o termo pertence . A Figura 3.4 mostra a representação gráfica dos dicionários.

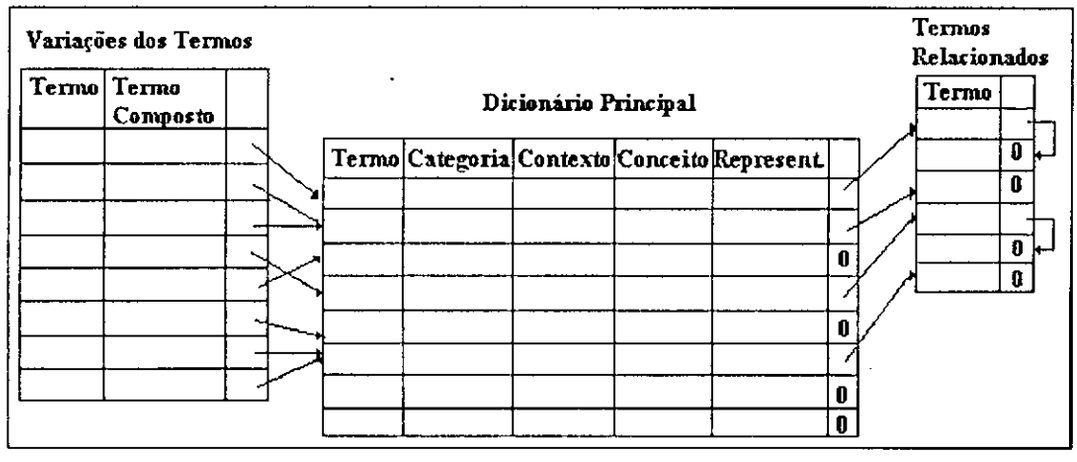


*Figura 3.4 – Representação gráfica da estrutura de um dicionário*

#### 3.2.1.4 Identificando conceitos relevantes

A análise de co-ocorrências entre pares de palavras é um meio para elucidar as estruturas das idéias e outros problemas representados em conjuntos adequados de documentos pelos seguintes princípios [Robredo, 1998]:

a) os autores de artigos científicos escolhem com cuidado os termos técnicos que utilizam;

b) quando diversos termos são utilizados no mesmo artigo, isso acontece, de fato, porque o autor reconhece ou supõe que existe algum tipo de relação não trivial entre seus referentes;

c) se um número significativo de autores reconhece o mesmo tipo de relacionamento entre determinados termos, pode-se admitir que esse relacionamento possui algum significado dentro da área da ciência considerada.

Neste sentido, calculamos um valor que expressa a força de ligação entre pares de conceitos, a partir da lista de termos extraídos dos documentos usando os trabalhos de Bruandet (Capítulo 2) como base. A medida da força de ligação será alterada para levar em consideração a freqüência com que o par de conceitos aparece apenas no documento sendo analisado e não no conjunto de documentos, isto para evitar os casos em que o par seja relevante para o conjunto, mas para determinados documentos, se calculado a parte ele não seria. Desta forma, garante-se que apenas os documentos relevantes para determinada consulta sejam recuperados.

Seja C o conjunto de conceitos extraídos do documento. É necessário que se defina uma medida que avalie a ligação contextual entre dois conceitos

quaisquer. Cada conceito pode ser identificado pela tupla <ND, NF, NC> onde ND é o número do documento, NF é o número da frase no documento, e NC é a posição do conceito na frase.

Representa-se a i-ésima ocorrência de um conceito x do vocabulário C, simbolizada por $w_i(x)$, da seguinte forma: $w_i(x) = <ND_i(x), NF_i(x), NC_i(x)>$. Para cada par de conceitos (x,y) que estejam no mesmo documento e na mesma frase, define-se uma distância d entre a i-ésima ocorrência de x e a j-ésima ocorrência de y, definida por:

$$\mathbf{d(w_i(x), w_j(y)) = \begin{cases} |NC_i(x) - NC_j(y)| & \text{se } \begin{cases} ND_i(x) = ND_j(y) \\ NF_i(x) = NF_j(y) \end{cases} \text{ e} \\ 0, \text{ caso contrário} \end{cases}}$$

A força de ligação entre dois termos, F, é definida como sendo o inverso da distância d da seguinte forma:

$$\mathbf{F(w_i(x), w_j(y)) = \begin{cases} \dfrac{1}{d(w_i(x), w_j(y))}, & \text{se } d(w_i(x), w_j(y)) \le t(x,y) \\ 0, \text{ caso contrário} \end{cases}}$$

onde

$\mathbf{d(w_i(x), w_j(y))}$ é a distância entre a ocorrência do par (x, y)

$\mathbf{t(x,y)}$ é um limite fixado experimentalmente para a distância entre dois conceitos.

O limite t(x,y) seleciona os conceitos cujas categorias gramaticais são interessantes para a qualificação do contexto do conceito, e é dado por:

t(x,y)= LIMITE[CAT(x), CAT(y)]

onde:

LIMITE é a distância máxima entre duas categorias gramaticais e

CAT é uma função que retorna a categoria gramatical de um conceito.

As categorias gramaticais e a distância máxima entre elas (t) podem ser modificadas em função de interesses específicos e/ou resultados obtidos.

A distância entre categorias gramaticais é definida por idioma, pois a distância entre um substantivo seguido por um adjetivo, além de ser diferente para um adjetivo seguido por um substantivo, não é a mesma, necessariamente, em inglês e português (Figura 3.5). Além disso, um substantivo-adjetivo, por exemplo, são considerados como relacionados semanticamente somente se aparecem a uma distância igual ou inferior a um determinado valor.

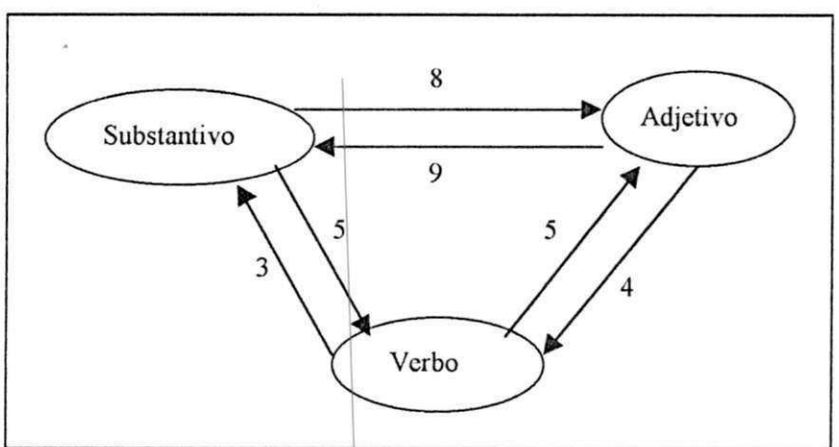


*Figura 3.5 – Exemplo do grafo das distâncias entre as categorias gramaticais para português.*

A medida de associação entre os conceitos x e y é dada por:

$$M(x,y) = \frac{b(x,y)}{f(x,y)} * k(x,y)$$

Onde:

$\mathbf{b(x,y) = \sum_i \sum_j F(wi(x), wj(y))}$ é o Somatório das Forças de Ligação F(x,y)) para todas as ocorrências i e j de x e y.

f(x,y) é a frequência do par(x,y) no documento

$\mathbf{k(x,y) = \frac{(f(x,y)-1)^n}{f(x,y)^n}}$, é um Fator de Correção onde n é um parâmetro inteiro definido por experimentação. Em [Bruandet, 1989b] o valor utilizado foi 2.

O fator de correção k é utilizado para eliminar os casos em que um par de conceitos (x,y) aparece uma única vez e são adjacentes numa mesma frase do conjunto de documentos ($0 \leq b(x,y)/f(x,y) \leq 1$). O fator k é zero quando a freqüência f(x,y) vale 1 e tende a 1 conforme f(x,y) aumenta. Através de k é possível atuar sobre a freqüência com a qual os pares de conceitos devem se relacionar. Um aumento do parâmetro n reforça as ligações muito freqüentes.

A Tabela 3.1 mostra o cálculo da Força de Ligação para um documento sobre o Orientação a Objetos, onde foram considerados para força mínima de ligação e o parâmetro *n* os valores 0,1 e 3, respectivamente. Apenas os pares que obtêm o valor da medida M maior que 0,1 (linhas em destaque) são considerados como relevantes para representar o documento.

| Termo 1 (t1) | Termo 2 (t2) | b(t1,t2) | f(t1,t2) | b(t1,t2)/f(t1,t2) | k | M (t1,t2) |
|---|---|---|---|---|---|---|
| **Orientado** | **Objeto** | **4,25** | **9** | **0,47** | **0,70** | **0,33** |
| **Análise** | **Orientado** | **3** | **3** | **1** | **0,29** | **0,29** |
| **Desenvolvimento** | **Software** | **2** | **4** | **0,5** | **0,42** | **0,21** |
| **Orientação** | **Objeto** | **2** | **4** | **0,5** | **0,42** | **0,21** |
| **Biblioteca** | **Classe** | **1,5** | **3** | **0,5** | **0,29** | **0,14** |
| **Mudança** | **Cultural** | **2** | **2** | **1** | **0,12** | **0,12** |
| **metodologia** | **Análise** | **1,16** | **3** | **0,39** | **0,29** | **0,11** |
| **conceito** | **Objeto** | **1,11** | **3** | **0,37** | **0,29** | **0,11** |
| **ambiente** | **Orientado** | **1,11** | **3** | **0,37** | **0,29** | **0,11** |
| análise | Objeto | 1 | 3 | 0,33 | 0,29 | 0,09 |
| banco | Dados | 1 | 2 | 0,5 | 0,12 | 0,06 |
| desenvolvimento | Objeto | 0,45 | 2 | 0,23 | 0,12 | 0,02 |

*Tabela 3.1: Exemplo do cálculo da Força de Ligação para um documento sobre orientação a objeto.*

### 3.2.1.5 Matriz conceito-conceito e Cliques

Os valores da medida de associação entre dois conceitos (medida M) são armazenados numa matriz conceito-conceito.

Através do parâmetro de força mínima de ligação é possível eliminar da matriz conceito-conceito as ligações mais fracas além de reduzir a quantidade de informações a serem armazenadas. A matriz conceito-conceito passa a ser uma matriz binária conceito-conceito, onde os conceitos que possuem força de ligação maior que a força mínima de ligação passam a receber o valor 1, e os demais são eliminados da matriz, ou seja, recebem o valor zero.

Gerando-se um grafo correspondente à matriz conceito-conceito, é possível extrair os subgrafos completos máximos, denominados **Cliques**. Os Cliques são subgrafos cujos nós estão todos conectados entre si (Figura 3.6). A idéia de se extrair cliques deve-se à dificuldade de analisar e interpretar as informações

diretamente na matriz. Pode-se dizer que um clique representa uma idéia contida no documento.

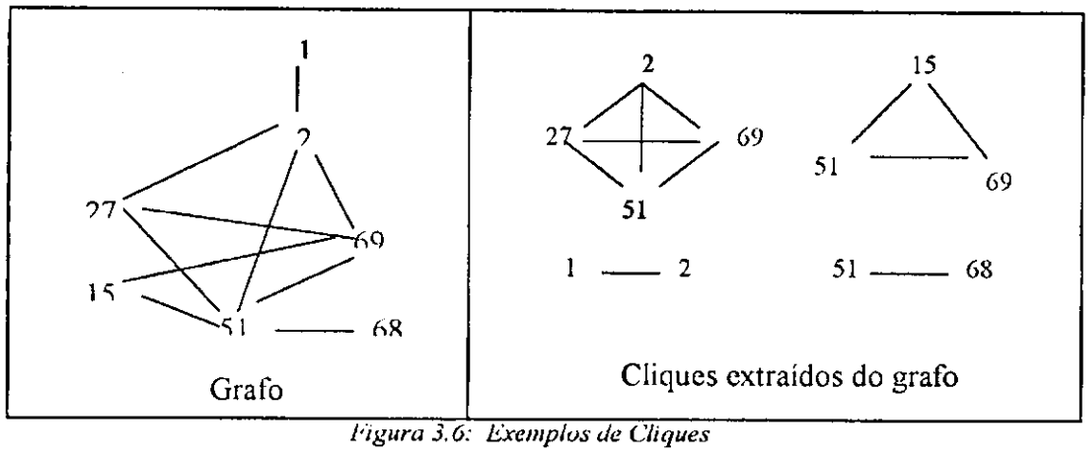


*Figura 3.6: Exemplos de Cliques*

Seja C={51, 69, 36, 37, 27, 2, 8, 5, 68, 28, 29, 1, 6, 7, 15 } o conjunto de conceitos extraídos de um documento. A seguir são mostrados a matriz conceito-conceito, que armazena os valores da medida M (associação entre cada par) (Figura 3.7), a matriz binária conceito-conceito (Figura 3.8), o grafo associado a matriz conceito-conceito (Figura 3.9), e os cliques extraídos da matriz binária conceito-conceito (Figura 3.10).

| **Conceito\Conceito** | 51 | 69 | 36 | 37 | 27 | 2 | 8 | 5 | 68 | 28 | 29 | 1 | 6 | 7 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 51 (Análise) | | 0,30 | | | | 0,09 | | | 0,12 | | | | | | |
| 69 (Orientado) | | | | | | 0,33 | | | | | | | | | 0,11 |
| 36 (Biblioteca) | | | | 0,15 | | | | | | | | | | | |
| 37 (Classe) | | | | | | | | | | | | | | | |
| 27 (Conceito) | | | | | | 0,11 | | | | | | | | | |
| 2 (Objeto) | | | | | | | | | | | | 0,21 | | | |
| 8 (Desenvolvimento) | | | | | | | | 0,21 | | | | | | | |
| 5 (Software) | | | | | | | | | | | | | | | |
| 68 (Metodologia) | | | | | | | | | | | | | | | |
| 28 (Mudança) | | | | | | | | | | | 0,13 | | | | |
| 29 (Cultural) | | | | | | | | | | | | | | | |
| 1 (Orientação) | | | | | | | | | | | | | | | |
| 6 (Tradicional) | | | | | | | | | | | | | | 0,13 | |
| 7 (Método) | | | | | | | | | | | | | | | |
| 15 (Ambiente) | | | | | | | | | | | | | | | |

*Figura 3.7: Matriz conceito-conceito representando a Medida M entre os pares de conceitos*

| Conceito\Conceito | 51 | 69 | 36 | 37 | 27 | 2 | 8 | 5 | 68 | 28 | 29 | 1 | 6 | 7 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 51 (Análise) | | 1 | | | | 0 | | | 1 | | | | | | |
| 69 (Orientado) | | | | | | 1 | | | | | | | | | 1 |
| 36 (Biblioteca) | | | | 1 | | | | | | | | | | | |
| 37 (Classe) | | | | | | | | | | | | | | | |
| 27 (Conceito) | | | | | | 1 | | | | | | | | | |
| 2 (Objeto) | | | | | | | | | | | | 1 | | | |
| 8 (Desenvolvimento) | | | | | | | | 1 | | | | | | | |
| 5 (Software) | | | | | | | | | | | | | | | |
| 68 (Metodologia) | | | | | | | | | | | | | | | |
| 28 (Mudança) | | | | | | | | | | | 1 | | | | |
| 29 (Cultural) | | | | | | | | | | | | | | | |
| 1 (Orientação) | | | | | | | | | | | | | | | |
| 6 (Tradicional) | | | | | | | | | | | | | | 1 | |
| 7 (Método) | | | | | | | | | | | | | | | |
| 15 (Ambiente) | | | | | | | | | | | | | | | |

*Figura 3.8: Matriz binária conceito- conceito*

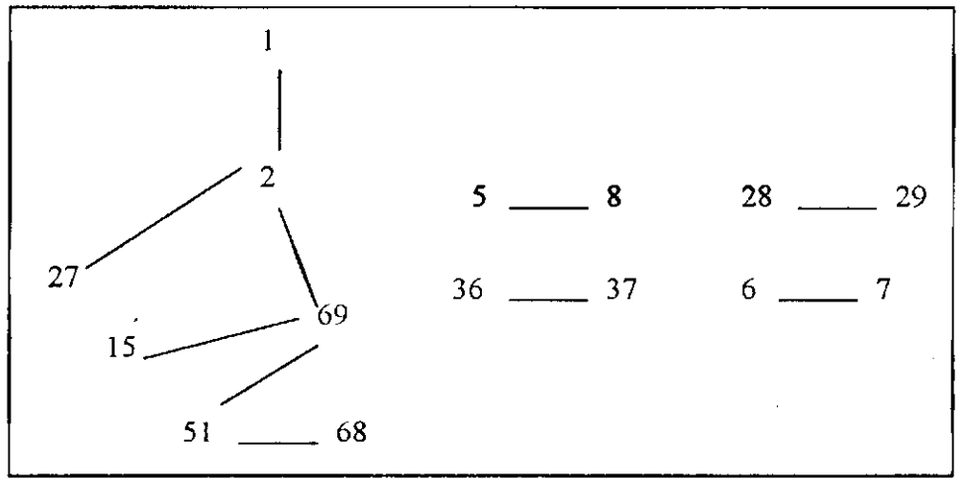


*Figura 3.9: Grafo da matriz binária conceito-conceito*

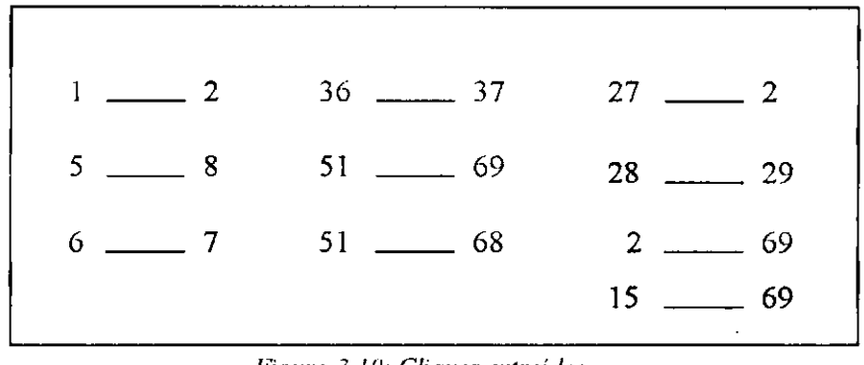


*Figura 3.10: Cliques extraídos*

### 3.2.2 GERAÇÃO DOS RETÂNGULOS ÓTIMOS E ATUALIZAÇÃO DO THESAURUS CONCEITO-CONCEITO

Os nós do thesaurus são retângulos ótimos obtidos através da decomposição retangular sobre a matriz binária conceito-conceito representada pela Figura 3.8.

Para cada documento analisado, gera-se os retângulos ótimos equivalentes que servem de entrada para o algoritmo incremental de PINTO [Pinto, 1997] que elimina retângulos redundantes da seguinte forma: quando um retângulo originado da interseção entre o novo retângulo a ser inserido e algum retângulo pertencente ao grafo é gerado, calcula-se o seu ganho, verificando se ele está ou não inserido no grafo. O grafo de retângulos (Figura 3.11) é construído em níveis hierárquicos relativos à cardinalidade do domínio (conjunto de termos) de cada retângulo.

A estrutura de dados[4] baseia-se numa lista vertical que contém os níveis de cardinalidade dos retângulos, permitindo assim estruturar o grafo hierarquicamente, o que facilita a inserção de um novo documento. Cada nó do grafo é composto por uma lista de: termos, documentos, pais, filhos e apontador para o próximo nó no mesmo nível de cardinalidade (retângulo vizinho).

O algoritmo de PINTO funciona da seguinte forma:

1. *Dado um novo documento, verifica-se a existência da cardinalidade de seus termos na lista de níveis de cardinalidade;*
2. *Caso não exista a cardinalidade na lista, insere-se o novo retângulo;*
3. *Caso exista a cardinalidade igual a do novo retângulo então*

[4] Apêndice B

*3.1. Se existir retângulo na lista de cardinalidade com termo igual ao termo do novo retângulo então insere-se o novo documento no retângulo existente e em seus filhos.*

*3.2. Se o novo documento não foi inserido, então insere-se.*

*4. Se ocorreu a inserção do novo retângulo então constrói-se a relação de ordem parcial do novo retângulo.*

*5. Se o termo está contido no supremo*

*então verifica-se as possíveis ligações com pais e filhos*

*Senão coloca-se o termo no supremo.*

*6. Se a lista de filhos do retângulo for vazia liga-se o retângulo ao Ínfimo;*

*7. Se a lista de pais do retângulo for vazia liga-se o retângulo ao supremo.*

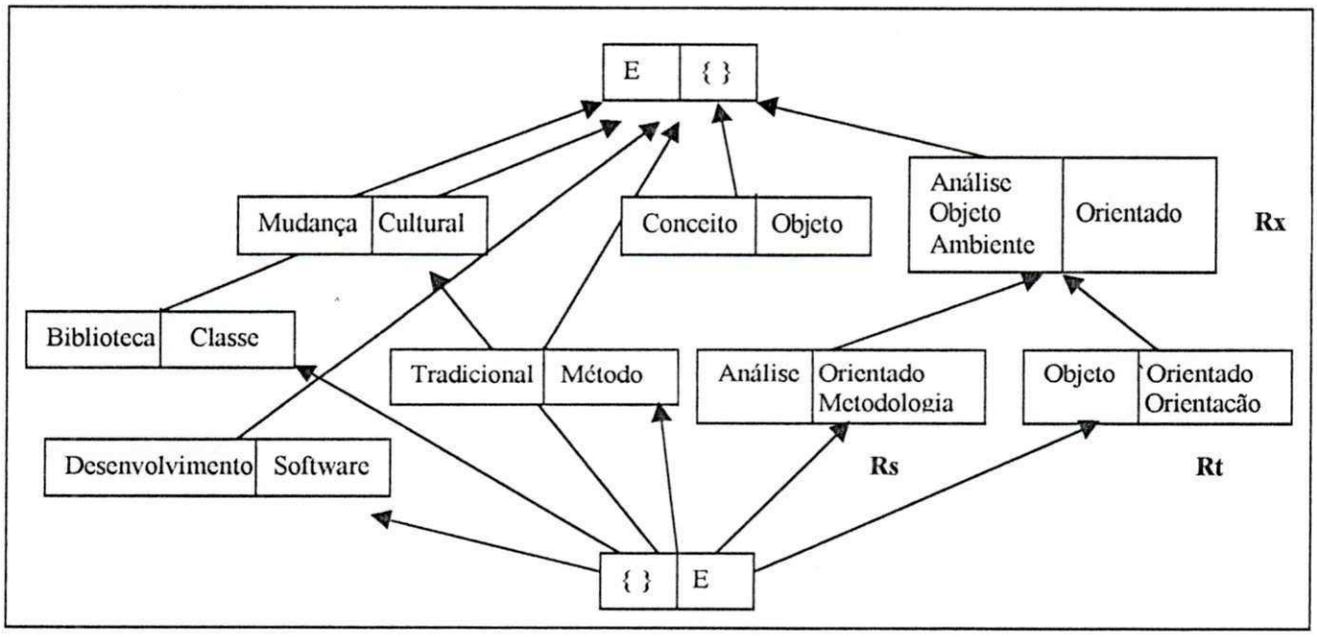


*Figura 3.11: Grafo de retângulos ótimos, em português, gerado a partir da matriz conceito-conceito.*

### 3.2.3 BASE DE DOCUMENTOS

A estrutura da base de documentos é representada através do Método de Decomposição Retangular [Gammoudi, 1993] a fim de manter a uniformidade e obter um melhor desempenho no sistema documental. Isto é, a base de documentos é representada por uma relação binária entre o conjunto de documentos D e o conjunto de conceitos C extraídos de D.

A organização hierárquica da base de documentos é realizada em três etapas:

1. Construção da matriz binária Conceito-Documento para representar a semântica da base;
2. Agrupamento dos documentos e seus descritores em forma de retângulos ótimos;
3. Geração do grafo de retângulos, que utiliza o mesmo algoritmo incremental de PINTO [Pinto, 1997].

A Figura 3.12 apresenta a matriz binária conceito-documento, para um conjunto de quatro documentos, sobre a qual aplica-se o algoritmo para geração do grafo representado na Figura 3.13.

| Conceito\Documento | 2 | 11 | 12 | 14 |
|---|---|---|---|---|
| 5 (Dados) | 1 | 1 | 1 | 1 |
| 19 (Técnica) | 1 | 1 | 1 | 1 |
| 22 (Tecnologia) | 1 | 1 | 1 | 1 |
| 32 (Conceito) | 1 | 1 | 1 | 1 |
| 37 (Metodologia) | 1 | 1 | 1 | 1 |
| 38 (Orientação) | 1 | 1 | 1 | |
| 40 (Objeto) | 1 | 1 | 1 | 1 |
| 44 (Linguagem) | 1 | 1 | 1 | 1 |
| 60 (Classe) | 1 | 1 | 1 | 1 |
| 72 (Modelo) | 1 | 1 | 1 | 1 |
| 79 (Software) | 1 | 1 | 1 | |
| 95 (Sistema) | 1 | 1 | 1 | 1 |
| 105 (Programação) | 1 | 1 | 1 | |
| 116 (Banco) | 1 | 1 | 1 | |
| 155 (Orientação a objeto) | 1 | 1 | 1 | |
| 158 (Biblioteca) | 1 | 1 | 1 | |
| 198 (Produtividade) | 1 | 1 | 1 | |
| 199 (Desenvolver) | 1 | 1 | | |
| 200 (Desenvolvimento) | 1 | 1 | 1 | |
| 203 (Aumento) | 1 | | 1 | |
| 212 (Tradicional) | | 1 | 1 | |
| 216 (Cliente) | | 1 | 1 | |
| 217 (Servidor) | | 1 | 1 | |
| 220 (Mudança) | 1 | 1 | 1 | |
| 221 (Cultural) | | 1 | 1 | |
| 222 (Ferramenta) | 1 | 1 | 1 | |
| 225 (Herança) | 1 | 1 | 1 | |
| 226 (Polimorfismo) | | 1 | 1 | |
| 227 (Relacional) | | 1 | 1 | |
| 229 (Estruturada) | | 1 | | |

*Figura 3.12: Matriz binária conceito-documento para um conjunto de quatro documentos*

Pela definição de retângulo genérico e específico (Capítulo 2), nota-se que a informação se repete no thesaurus. Por exemplo, na Figura 3.11 o retângulo ótimo Rs, cujo domínio é composto pelo termo "análise" é mais genérico do que o retângulo Rx. Por isso, o termo "análise" também está presente no retângulo Rx.

Para diminuir a quantidade de informação a ser armazenada, escolhe-se para cada retângulo o seu representante que é um subconjunto ou do seu domínio

ou do seu codomínio, ou seja, serão extraídos de cada retângulo os termos que aparecem em seus ancestrais diretos.

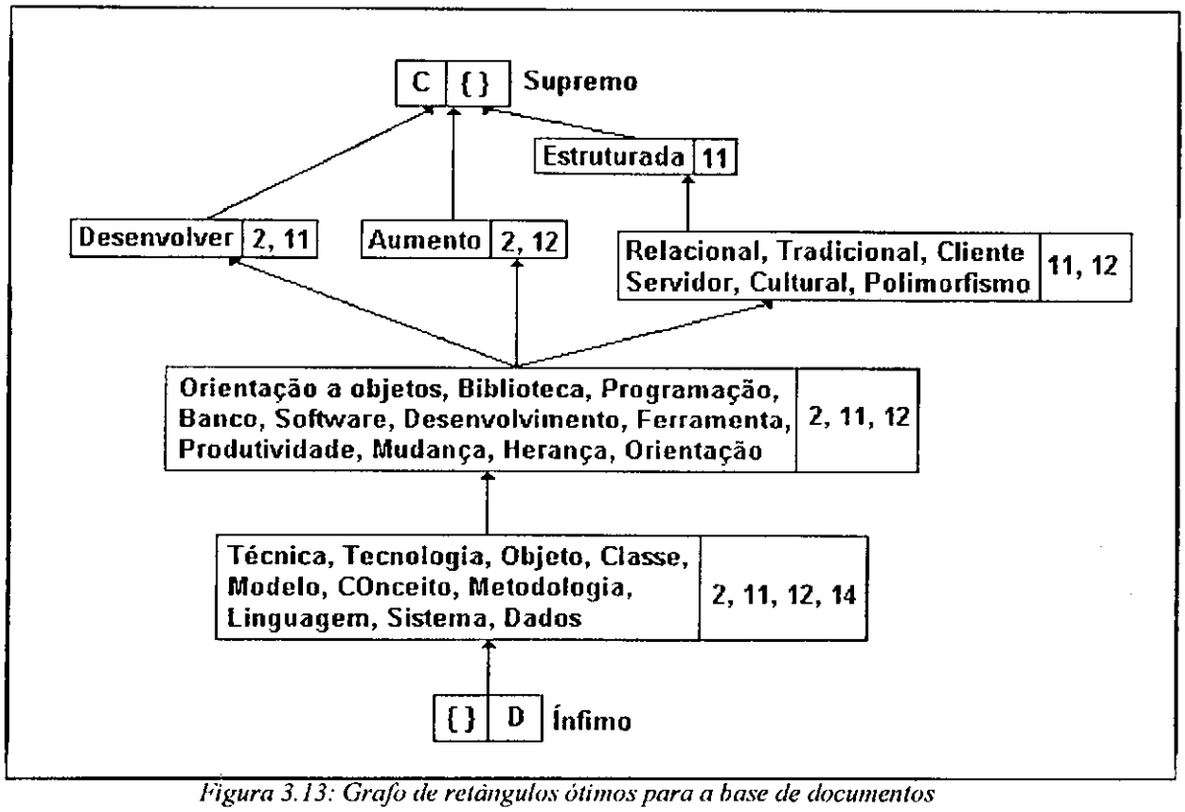


*Figura 3.13: Grafo de retângulos ótimos para a base de documentos*

***Definição 3.1: Representante de um retângulo***

---

*Seja o retângulo ótimo* $R_i = (A_i, B_i) \in R_{ótimo}$. *O representante de* $R_i$ *é:*

$$Ai, \text{ se } Card(A_i) \leq Card(B_i)$$

$$Bi, \text{ se } Card(B_i) < Card(A_i)$$

---

O representante de um retângulo contém os termos de maior conectividade. A *conectividade* de um termo é o número de ligações que ele possui com os outros termos do mesmo retângulo. Quanto mais um termo é conexo, mas ele veicula semântica e permite assim um maior número de termos [Ferneda, 1997].

Dois casos particulares podem acontecer no processo de simplificação [Pinto, 1997]:

1. Após o processo de simplificação de um retângulo Rx seu nó correspondente Nx for vazio. Neste caso faz-se Nx=Rx.Termo.

2. Após o processo de simplificação, um nó Nx ficar com mais de um elemento, tais termos são pseudo-sinônimos.

A Figura 3.14 mostra que o nó Nx, é o resultado da simplificação do retângulo Rx da Figura 3.11, isto é:

$$Nx = Rx.termo - \{ \{Rx \cap Rs\} \cup \{Rx \cap Rt\} \} =$$

$$= \{Análise, Objeto, Ambiente\} - \{ \{ Análise\} \cup \{Objeto\} \}$$

$$= Ambiente.$$

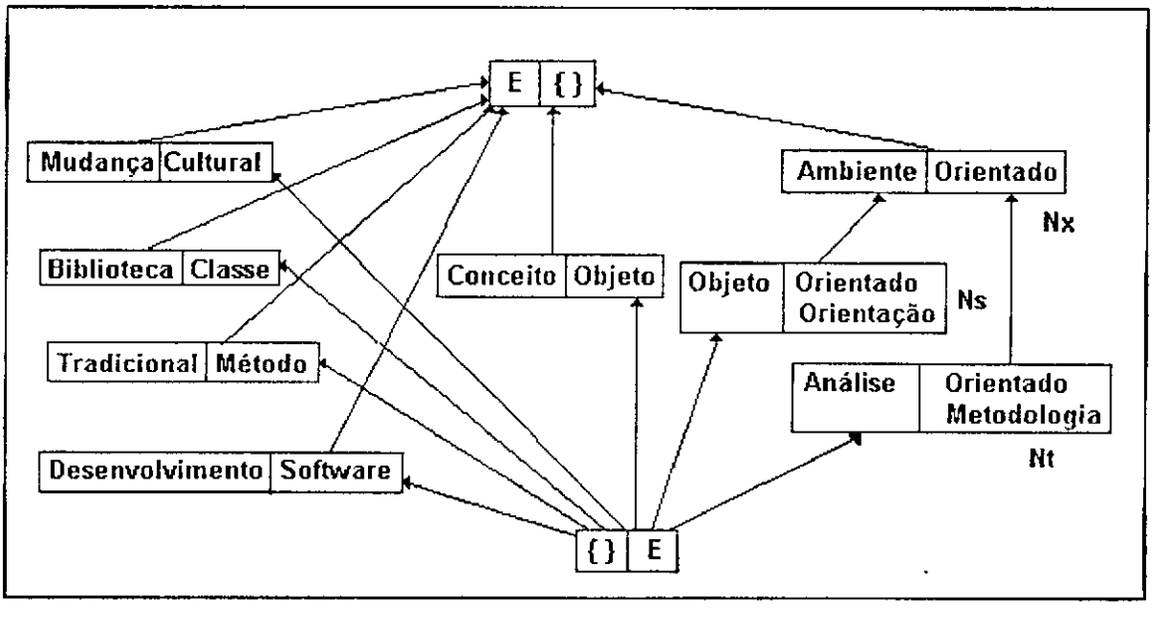


*Figura 3.14: Grafo de retângulos ótimos simplificado em português*

# 4 PROTÓTIPO PARA CONSTRUÇÃO SEMI-AUTOMÁTICA DE THESAURUS RETANGULAR MULTILÍNGUE - SISMULT

Este capítulo tem como objetivo apresentar um protótipo para a construção semi-automática de thesaurus retangular multilíngue utilizando o método proposto no Capítulo 3. Além disso, este protótipo pode vir a ser um produto acabado uma vez que existe no Departamento de Sistemas e Computação, Universidade Federal da Paraíba -Campus II, um projeto denominado *SICRET – Sistema Inteligente de Cadastro, Recuperação e Empréstimo de Títulos*, cujo objetivo é atender às bibliotecas atuais tanto na catalogação de documentos em meios normais, como o papel, como a indexação de documentos digitais. Denominamos este protótipo como *SISMULT – Sistema de Indexação Semi-automática Multilíngüe.*

## 4.1 INDEXAÇÃO DE DOCUMENTOS

O principal objetivo do SISMULT é a construção semi-automática de um thesaurus retangular multilíngüe, a partir de documentos eletrônicos em ou mais idiomas, em três etapas (capítulo 3): extração de termos relevantes dos documentos, geração dos retângulos ótimos a partir das matrizes geradas pelo processo de extração, e geração do thesaurus retangular multilíngüe utilizando um algoritmo incremental para inserir os retângulos ótimos.

Identifica-se dois tipos de usuários para o sistema: o biblioteconomista, ou usuário responsável pelo cadastro e manutenção do acervo, que acompanharia o processo de construção/atualização do thesaurus retangular

multilíngüe; e o usuário da biblioteca, cujo interesse é selecionar documentos que atendam à sua consulta. Logo, é importante que o protótipo disponibilize um módulo para recuperação de informação para que o usuário possa verificar os resultados do método através de consultas, uma vez que nas etapas de geração de retângulos ótimos e de inserção destes no thesaurus não há interação com o usuário.

O sistema foi desenvolvido na linguagem de programação Delphi 3.0 (maiores referências sobre esta linguagem podem ser encontradas em [Cantú, 1996]), em ambiente Windows 95, no periodo de quatro meses, com a colaboração de um programador.

Os formatos permitidos para os documentos eletrônicos, atualmente, são o formato texto (extensão txt) e o formato rich text format (extensão rtf), nos quatro idiomas considerados: português, inglês, francês e italiano.

O SISMULT permite além da construção/atualização do thesaurus e da consulta ao thesaurus, a edição de dicionários – dicionário principal e dicionário de stopwords, ambos por idioma; configuração de parâmetros utilizados no método de extração de termos (capítulo 3), como distância entre categorias gramaticais e valor mínimo da força de ligação; e disponibiliza ao usuário uma lista de palavras não encontradas nos dicionários que pode ser usada na manutenção dos próprios dicionários.

O sistema permite a definição de um ou mais thesauri através da definição dos documentos que serão utilizados na sua construção.

A Figura 4.1 apresenta a janela onde o usuário define os seus thesauri. A janela permite ao usuário adicionar documentos, alterar informações de um

documento, além de apresentar o conteúdo do documento em foco. Além disso, é mostrada a situação de cada documento, isto é, se já foi processado (inserido no grafo retangular), se não existe na localização indicada e se ainda não foi processado.

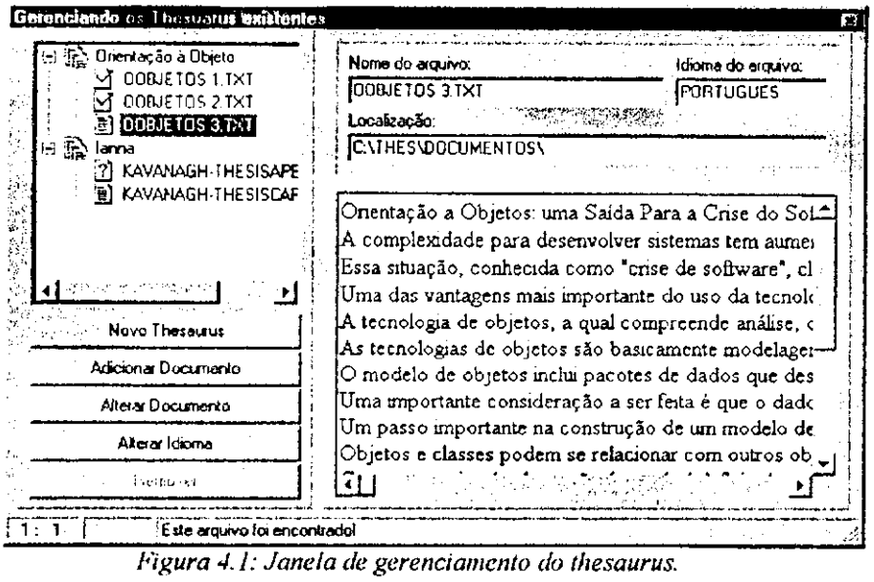


*Figura 4.1: Janela de gerenciamento do thesaurus.*

Após a definição dos thesauri o processo de extração de termos pode ser iniciado, sendo necessário para tanto apenas que o usuário ative um dos thesauri existentes.

A extração de termos é a primeira fase a ser executada pelo sistema a fim de construir o thesaurus abstrato. Na Figura 4.2, tela de extração de termos, identifica-se o thesaurus que está sendo processado, o conteúdo do texto em análise, e o progresso da execução da extração.

A indexação dos termos do conjunto de documentos do thesaurus ativo é feita de uma única vez e semi-automaticamente, isto é, para as palavras ambíguas encontradas nos documentos o sistema interage com o usuário para que este escolha o contexto adequado.

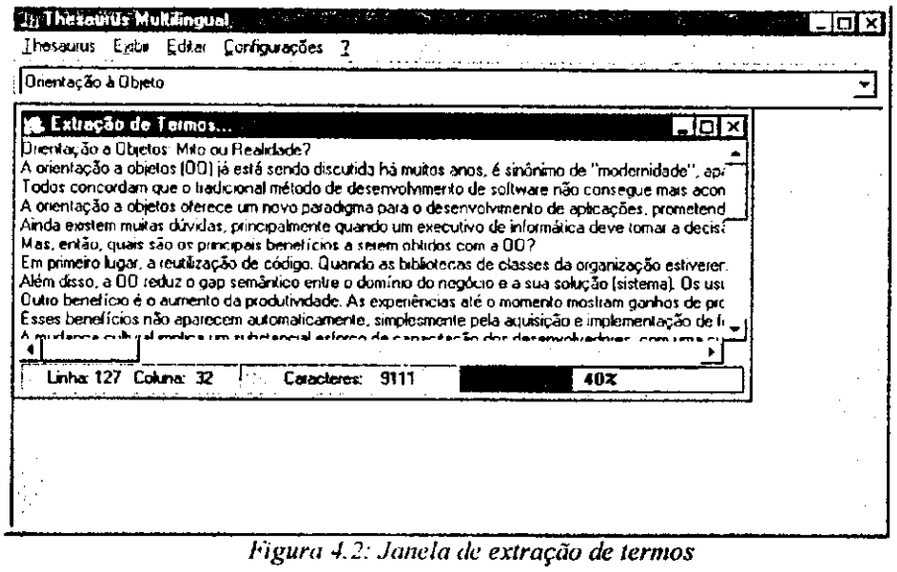


*Figura 4.2: Janela de extração de termos*

Para auxiliar o usuário na escolha do contexto apropriado criamos uma janela denominada 'resolução de conflitos' (Figura 4.3) onde é possível visualizar o termo ambíguo e seus contextos, além do parágrafo onde o termo foi encontrado. Para cada contexto são apresentados os termos sinônimos e os termos relacionados, se existirem, e a categoria gramatical.

Implementamos a opção 'usar sempre', que funciona por documento, para tornar o processo de extração automático no sentido de utilizar sempre um determinado contexto, a fim de evitar que esta ação se torne enfadonha e também garantir que será sempre usado para um termo o mesmo conceito. A opção 'ignorar sempre' foi implementada para que o usuário tenha a liberdade de não escolher nenhum contexto dos que lhe foi apresentado pelo sistema, ou seja, nenhum dos contextos é adequado, logo o termo identificado não será indexado.

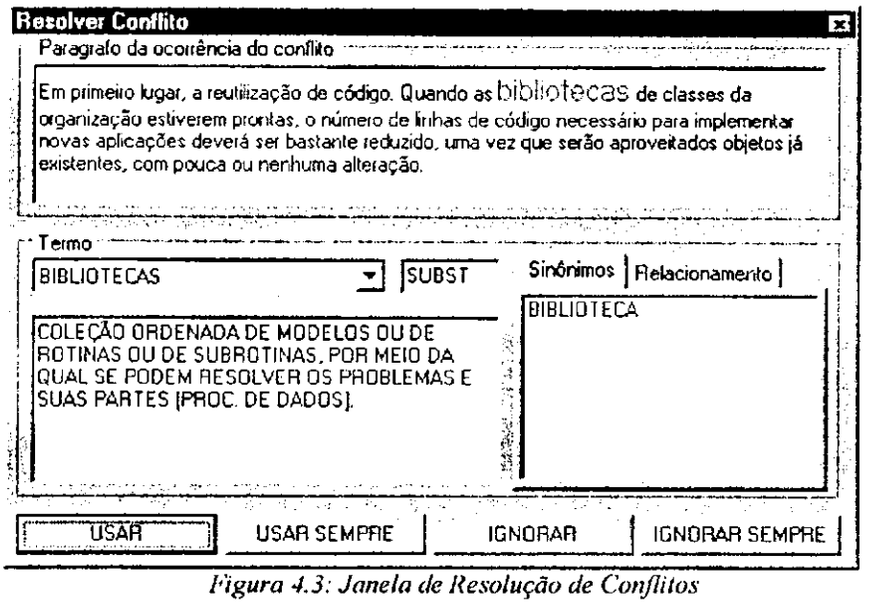


*Figura 4.3: Janela de Resolução de Conflitos*

A construção dos vários dicionários utilizados na extração de termos se deu manualmente, assim como a construção de classes de palavras sinônimas, às quais é atribuído um único conceito. Para tanto, foi criado um formulário para facilitar as entradas dos termos, ou seja, os dicionários de variações de termos e o de termos correlacionados são construídos automaticamente a partir da inserção de termos no dicionário principal.

A Figura 4.4 apresenta o formulário para inserção de novos termos ao dicionário principal. Antes da inserção do novo termo é necessário que se escolha o idioma do termo a ser inserido para que os dicionários equivalentes sejam ativados.

São três as categorias gramaticais consideradas neste trabalho: substantivo, adjetivo e verbo. Dependendo da categoria escolhida para o novo termo, novas guias são habilitadas. Por exemplo, a Figura 4.4 mostra que o termo a ser inserido é 'Comunicar', categoria gramatical 'VERBO'. Logo é necessário habilitar um formulário para a conjugação do verbo (Figura 4.5).

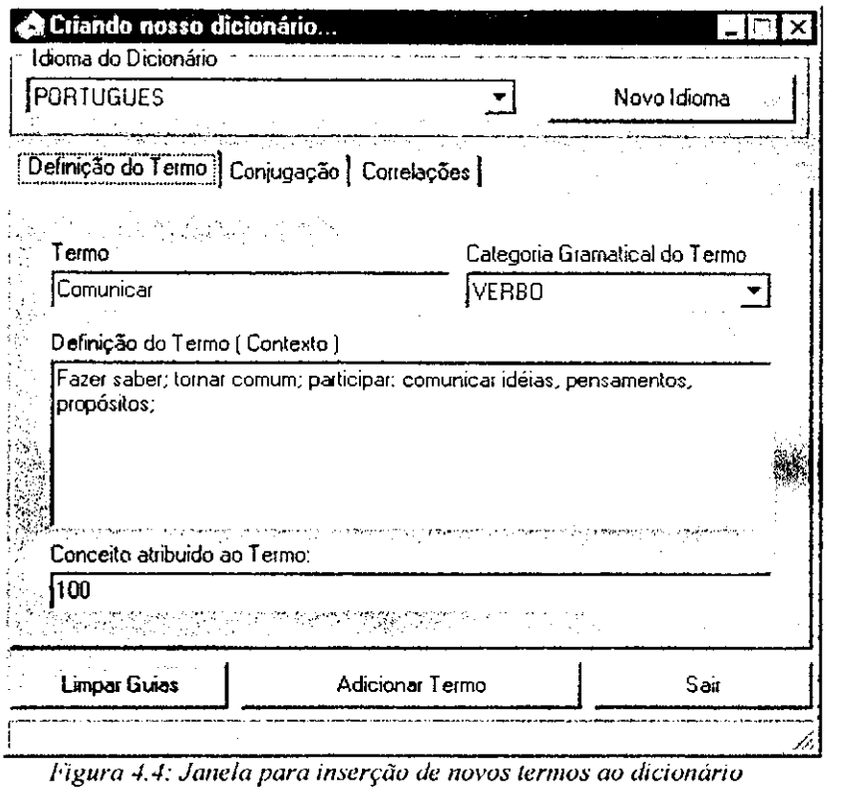


*Figura 4.4: Janela para inserção de novos termos ao dicionário*

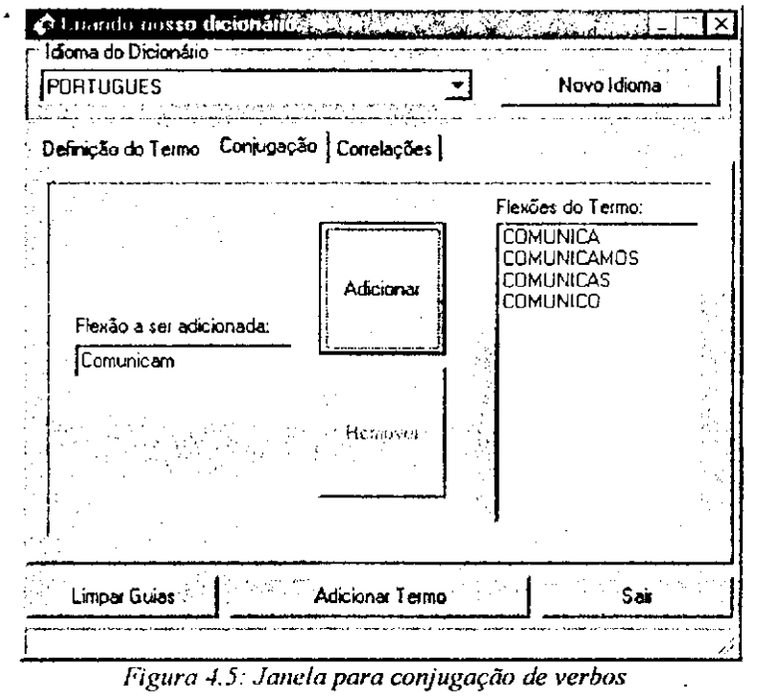


*Figura 4.5: Janela para conjugação de verbos*

No caso do termo ser substantivo ou adjetivo, as flexões do termo podem ser inseridas: plural, singular, termo feminino, termo masculino, aumentativo e diminutivo (Figura 4.6).

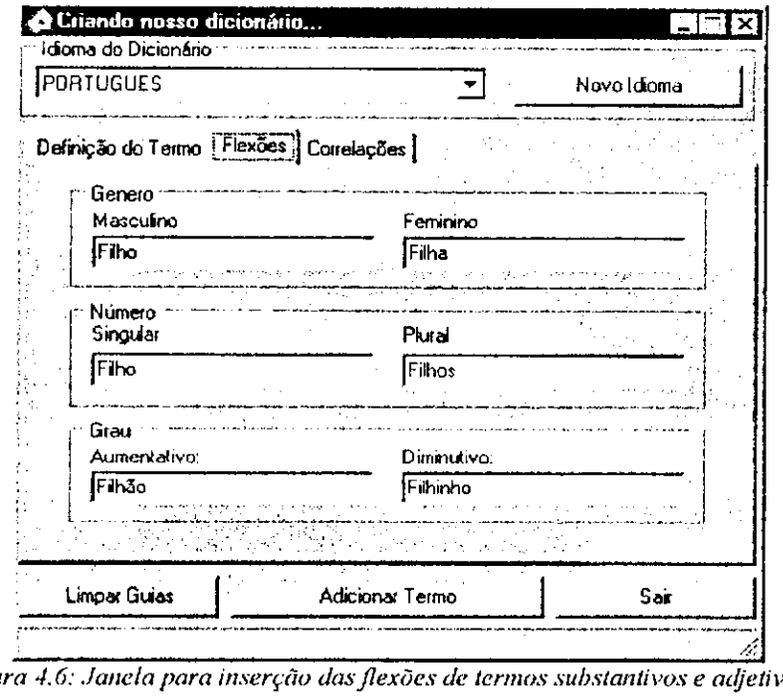


*Figura 4.6: Janela para inserção das flexões de termos substantivos e adjetivos.*

As stopwords são palavras sem valor de indexação, como pronomes, conjunção, preposição, etc. Para cada idioma é construído inicialmente um dicionário de stopwords que pode ser atualizado após a extração de conceitos de um ou mais documentos, a partir da lista de palavras não encontradas gerada pelo sistema.

A Figura 4.7 mostra uma lista de stopwords para o idioma português e no lado direito apresenta a lista de palavras não encontradas para um conjunto de documentos. É possível, além da inserção de novas stopwords a partir da lista de palavras não encontradas, a remoção de algumas. Isto porque, a lista de stopwords pode ser útil também caso o usuário deseje nunca indexar uma determinada palavra. A lista de palavras não encontradas pode ser usada também para atualizar os

dicionários do sistema, mas esta opção não está implementada nesta versão do SISMULT.

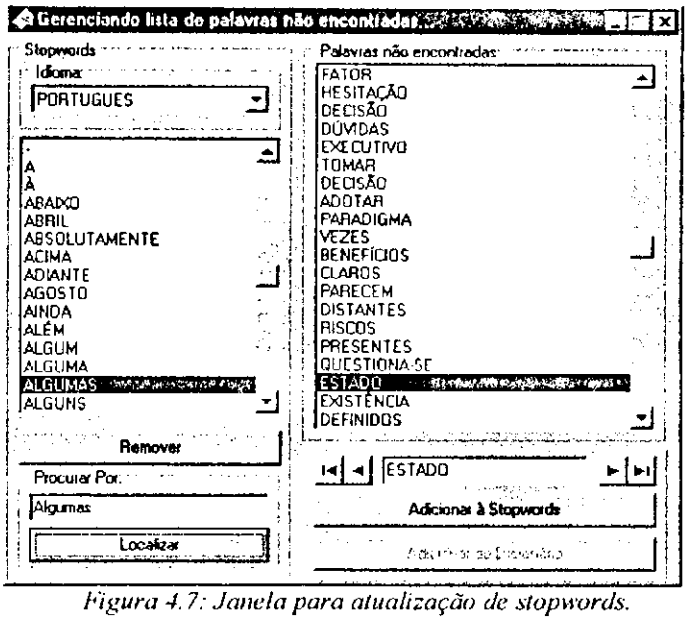


*Figura 4.7: Janela para atualização de stopwords.*

Finalizada a extração dos conceitos, cada documento será processado individualmente para ser inserido no grafo retangular. A construção do grafo retangular implica na construção da matriz conceito-conceito, extração de cliques, geração da matriz conceito-documento, geração do relacionamento conceito-documento e por fim da inserção dos retângulos obtidos no grafo. A Figura 4.8 informa ao usuário qual etapa está sendo executada num determinado momento.

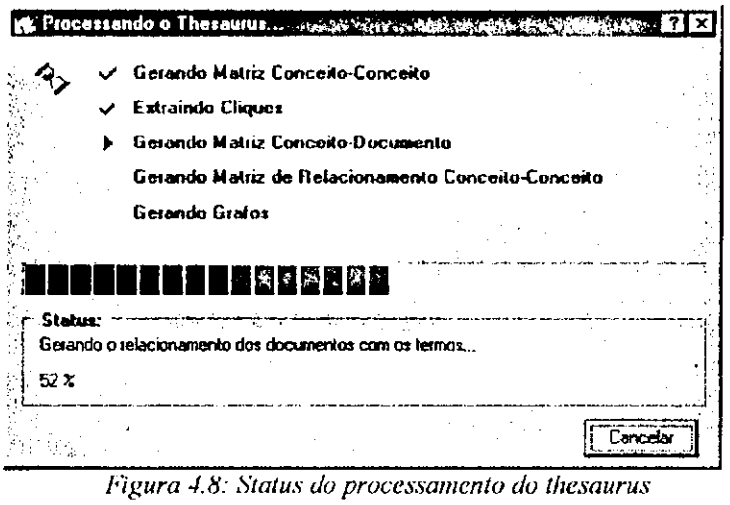


*Figura 4.8: Status do processamento do thesaurus*

## 4.2 CONSULTAS

Finalizado o processo de construção do thesaurus, a opção de consulta se torna disponível ao usuário. A opção de consulta ao thesaurus tem como objetivo apresentar ao usuário um conjunto de documentos que satisfaçam sua consulta.

Em geral, uma consulta pode ser uma expressão booleana de palavras-chave, uma expressão em linguagem natural ou ainda pode ser construída a partir de uma seqüência de seleções de menus ou nós em um grafo. Na maioria dos sistemas de recuperação atuais, por exemplo como os browsers Yahoo, Infoseek, Altavista, etc., disponíveis na Internet, os usuários formulam suas consultas através de expressões, que contêm palavras (que supostamente indexarão os documentos de seu interesse) combinadas pelos operadores AND e OR [Pinto, 1997]. No entanto, este tipo de interface apresenta três problemas:

1. em geral, o usuário fica impossibilitado de construir consultas mais complexas, a fim de recuperar informações relevantes, por não dominar a álgebra de Boole;

2. em resposta à sua consulta o usuário recebe uma lista grande de documentos, mas nenhuma informação sobre o grau de relevância deles;

3. como dificilmente o usuário tem conhecimento sobre os termos utilizados no banco de dados documental, ele não consegue utilizar em suas consultas todas as palavras classificadas como termos de indexação, não recuperando todos os documentos relevantes.

Uma maneira de melhorar a recuperação da informação é possibilitar que o usuário tenha acesso a um thesaurus [Lewis,1996], permitindo que seus termos sejam utilizados para incrementar o processo de elaboração de consultas [Pinto, 1997].

A opção de consulta no SISMULT foi implementada de forma que o usuário não precise ter conhecimento sobre álgebra de Boole, pois o usuário pode visualizar os retângulos do thesaurus simplificado, no idioma desejado, utilizando na sua consulta apenas os termos indexados que são apresentados na tela.

Como o retângulo escolhido para a recuperação de documentos pode conter mais conceitos além do especificado na consulta, estes conceitos poderão ser uma informação útil ao usuário para caracterizar os documentos de sua consulta.

A interface de consulta é composta dos seguintes objetos:

**'Termos'**- Apresenta a lista, em ordem alfabética, dos termos contidos no thesaurus simplificado escolhido pelo usuário;

**'Genérico'** – Apresenta a lista, ordenada pelo grau de generalidade (Capítulo 2) dos termos contidos no thesaurus que são ancestrais diretos (pais) do termo em foco em 'Termos';

**'Específico'** - Apresenta a lista, ordenada pelo grau de especificidade (Capítulo 2) dos termos contidos no thesaurus que são descendentes diretos (filhos) do termo em foco em 'Termos';

**'Sinônimo'** - Apresenta a lista dos sinônimos do termo em foco em 'Termos'. Os termos sinônimos são aqueles que possuem o mesmo conceito, ou seja, pertencem à mesma classe conceitual e são identificados nos dicionários;

**'Pseudo-sinônimo'** - Apresenta a lista dos termos, que também pertencem ao nó do termo em foco, que foram eliminados no processo de simplificação do thesaurus;

**'Vizinhos'** - Apresenta uma lista, ordenada pelo grau de vizinhança (Capítulo 2) dos termos contidos no thesaurus que tenham a mesma cardinalidade do termo em foco em 'Termos';

**'Documentos'** – Apresenta os documentos que contêm o termo em foco, isto é, nome do documento, idioma e endereço, podendo inclusive serem visualizados.

Para cada nó do grafo são exibidos os retângulos específicos e genéricos que são ancestrais e descendentes diretos do termo em foco, respectivamente, o que permite ao usuário modificar sua consulta. A Figura 4.9, tela de consulta do SISMULT, apresenta um exemplo de consulta onde o termo pesquisado é 'técnica' no idioma 'português'. Nota-se que através dos objetos da interface é possível percorrer o grafo de retângulos, e ainda que os operadores *and* e *or* são desnecessários uma vez que termos sinônimos (operador *or*) e termos específicos (operador *and*) são apresentados na tela. Qualquer termo em qualquer objeto da tela pode servir como termo de consulta.

Esta interface permite rápido acesso aos nós do thesaurus e a navegação ocorre de forma "limpa", pois apenas os termos do thesaurus relacionados ao termo em foco, consequentemente ao thesaurus de interesse do usuário, são apresentados na tela durante a navegação.

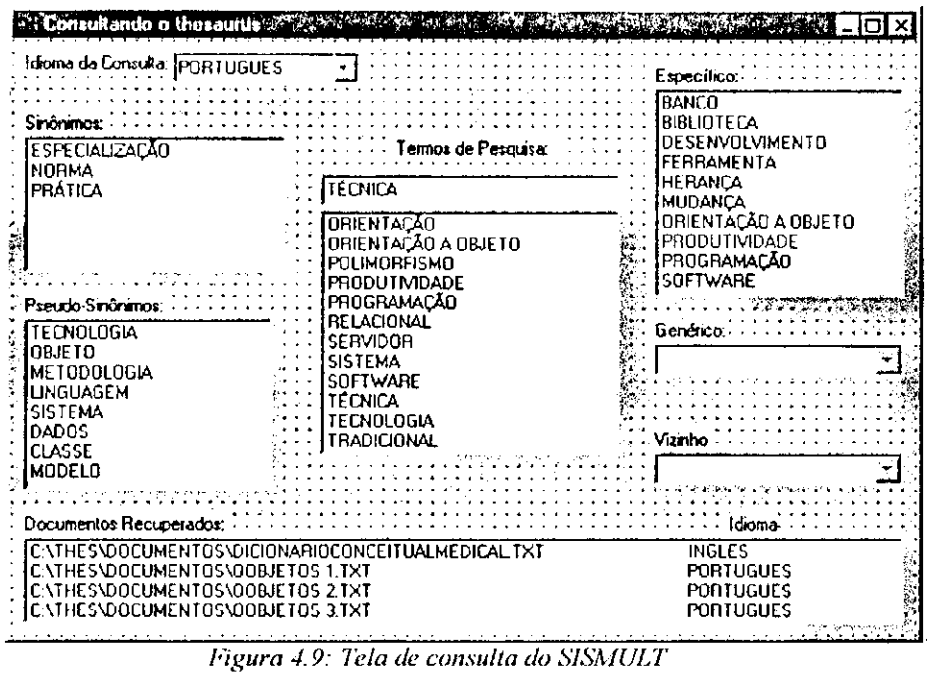


*Figura 4.9: Tela de consulta do SISMULT*

## 4.3 EXPERIMENTAÇÃO

Foi feita uma experimentação da fase de extração de termos num conjunto de nove documentos (Tabela 4.1), em italiano e português, contendo um total de 15.290 palavras, e outra num conjunto de três documentos (Tabela 4.2), estes todos em inglês, com um total de 18.209 palavras.

A Tabela 4.2 apresenta um conjunto de documentos que tratam do mesmo assunto, enquanto que na Tabela 4.1 os documentos são de assuntos diversos. Podemos afirmar que os dicionários são de grande importância para esta fase, pois quanto mais completo maior o número de palavras encontradas, o que eleva o número de conceitos distintos representantes dos documentos. Isso pode ser notado pelo fato de que na Tabela 2, os documentos tratam sobre análise textual, tema

coincidente com o deste trabalho e para o qual o dicionário foi construído. Logo, o número de pares conceito-documento é maior do que os da tabela 4.1.

O tempo de processamento para ambas as experimentações (tabela 4.1 e tabela 4.2) foi de aproximadamente 30 e 35 minutos, respectivamente, utilizando um Pentium de 100 Mhz com 16 Mb de Ram. É evidente que este tempo de processamento precisa ser melhorado, o que implica em modificações nos algoritmos utilizados nesta fase.

| Documento | Idioma | Total de Palavras | Palavras Encon-tradas | Pares matriz Conceito-Documento | Pares Matriz Conceito-Conceito |
|---|---|---|---|---|---|
| 1. Traducaomaquina-ita.rtf | Italiano | 1.163 | 108 | 8 | 6 |
| 2. Tecnologiasestadofinito-ita.rtf | Italiano | 1.583 | 114 | 7 | 6 |
| 3. Ocr-ita.txt | Italiano | 1.607 | 174 | 16 | 11 |
| 4. Introdução.rtf | Português | 1207 | 182 | 15 | 18 |
| 5. Lexicons-ita.txt | Italiano | 1.432 | 204 | 23 | 40 |
| 6. OObjetos 1.txt | Português | 1.348 | 263 | 31 | 43 |
| 7. OObjetos 3.txt | Português | 1.574 | 286 | 43 | 72 |
| 8. Multilingualsppech-ita.txt | Italiano | 3.199 | 292 | 23 | 46 |
| 9. OObjetos 2.txt | Português | 2.177 | 365 | 20 | 26 |
| **T O T A L** | | **15.290** | **1988** | **186** | **268** |

*Tabela 4.1 – Conjunto de documentos processado pelo SISMULT*

| Documento | Idioma | Total de Palavras | Palavras Encon-tradas | Pares matriz Conceito-Documento | Pares matriz Conceito-Conceito |
|---|---|---|---|---|---|
| 1. Kavanagh-thesisApendice.txt | Inglês | 472 | 15 | 0 | 0 |
| 2. Kavanagh-thesiscap2.txt | Inglês | 5.675 | 777 | 66 | 222 |
| 3. Kavanagh-thesiscap3.txt | Inglês | 12.062 | 1115 | 52 | 184 |
| **T O T A L** | | **18.209** | **1907** | **118** | **406** |

*Tabela 4.2 – Conjunto de documentos processado pelo SISMULT*

Após a extração de termos, as demais etapas (geração da matriz conceito-conceito, geração de cliques e geração da matriz conceito-documento) são executadas por documento e rapidamente.

# 5 CONSIDERAÇÕES FINAIS

Este capítulo tem o objetivo de tecer comentários a respeito do método e do protótipo SISMULT para construção semi-automática de thesaurus retangular multilíngüe, além de sugerir alguns trabalhos adicionais que poderiam ser desenvolvidos para aperfeiçoar o protótipo.

- O método desenvolvido amplia os estudos de thesauri unilíngües para aplicações em diversos idiomas utilizando documentos eletrônicos em linguagem natural, proporcionando uma indexação e recuperação de informações satisfatórias;

- a indexação semi-automática, apesar de demandar mais tempo de processamento do que a automática, consegue eliminar ambigüidades e permite ao usuário escolher o contexto correto a ser usado;

- a análise de co-ocorrência usada na construção do SISMULT pode ser útil em outras aplicações, como, por exemplo, na aquisição de documentos numa biblioteca através da comparação do conteúdo temático dos documentos (livros, periódicos, anais, etc.) com os programas curriculares;

- Observou-se que o SISMULT na fase de extração de termos apresenta um baixo tempo de resposta, sendo necessária portanto uma revisão dos algoritmos utilizados e a realização de testes com grande volume de dados;

- O protótipo desenvolvido permite uma atualização contínua dos thesauri existentes com novos documentos, em diversos idiomas, e a realização de consultas multilíngües, além de permitir o acréscimo de novos idiomas, sendo portanto uma solução para a paráfrase "como recuperar documentos que contêm

expressões que não casam exatamente com aquelas estabelecidas na consulta? [Fluhr, 1995]";

- Como os dicionários utilizados no protótipo foram elaborados com propósito único de aplicação neste trabalho, já que os existentes não atenderam aos requisitos, observou-se que para uso genérico eles são limitados embora tenham validado o propósito de execução do protótipo SISMULT.

Alguns trabalhos adicionais interessantes podem ser desenvolvidos tais como:

- expandir os formatos permitidos de textos eletrônicos, pois atualmente apenas os formatos texto (txt) e rich text format (rtf) são permitidos;
- incorporar um processo de digitalização de documentos e reconhecimento de caracteres;
- viabilizar a remoção de documentos dos thesauri sem que haja a necessidade de reprocessá-los;
- incluir outros alfabetos para que seja possível o tratamento de idiomas como chinês e japonês;
- permitir a atualização automática dos dicionários a partir da lista de palavras não-encontradas gerada pelo SISMULT;

## REFERÊNCIAS BIBLIOGRÁFICAS


[Aitchison, 1979] **AITCHISON, Jean.** *Manual para construção de tesauros* / Jean Aitchison e Alan Gilchrist; tradução de "Thesaurus Construction: a practical manual" por Helena Medeiros Pereira Braga. – Rio de Janeiro: BNG/Brasilart, 1979.

[Attar, 1977] **ATTAR, R. & S. Fraenkel.** *Local Feedback in Full-Text Retrieval System.* Journal of the ACM, v. 24, n. 3, Julho, 1977, p. 397-417.

[Bruandet, 1980a] **BRUANDET, Marie-France.** *A Conceptual Framework for Automatic and Dynamic Thesaurus Updating in Information Retrieval Systems.* COLING'80. Proceedings of the 8th International Conference on Computational Linguistic. Setembro/Outubro, 1980.

[Bruandet, 1980b] **BRUANDET, Marie-France.** *A Propos de La Construction Automatique d'un Thésaurus dans un Système de Recherche d'Information (Système Documentaire).* IMAG. Rapport de Recherche (Relatório Técnico) nº 229. Novembro, 1980.

[Bruandet, 1981] **BRUANDET, Marie-France.** *Notion de Concept pour la Construction Automatique d'un Thésaurus Evolutif.* AFCET Informatique. Actes du Congrès de l'Afcet. Editions Hommes et Techniques.18-20 de Novembro, 1981.

[Bruandet, 1982] **BRUANDET, Marie-France, CHIARAMELLA, Y., KERKOUBA, D.,** *Méthodes d'Indexation Automatique de Documentations Techniques dans le Cadre d'un Atelier de Logiciel.* Journées d'études CONCERTO. Perros-Guirec 16-17 Dezembro, 1982.

[Bruandet, 1982a] **BRUANDET, Marie-France.** *Concept Notion of Automatic and Dynamic Thesaurus Updating.* Conference Proceedings International Conference on System Documentation – SIGDOC. Carson, Califórnia. Janeiro, 1982.

[Bruandet, 1985] **BRUANDET, Marie-France.** *Modèle Partiel de Connaissances pour un Système de Recherche d'Information.* Recherche d'Informations Assistée par Ordinateur – RIAO'85. Grenoble, France. 18-20 de Março , 1985.

[Bruandet, 1989a] **BRUANDET, Marie-France.** *Outline of a Knowledge-Base Model for an Intelligent Information Retrieval System.* Information Processing & Management. v. 25, n. 1, 1989, p. 89-115.

[Bruandet, 1989b] **BRUANDET, Marie-France.** *Construction Automatique d'une Base de Connaissances du Domaine dans un Système de Recherche d'Information.* Document fourni pour la soutenance du Diplôme d'Habilitation à Diriger des Recherches de L'Université Joseph Fourier de Grenoble. 13 de Março, 1989.

[Bussmann, 1995] **BUSSMANN, José Eduardo Carvalho.** *Bart- Uma Biblioteca Orientada a Objetos de Apoio à Recuperação Textual.* Dissertação de Mestrado. Universidade Federal da Paraíba. Dezembro, 1995.

[Cantú, 1996] **CANTÚ, Marco.** *Dominando o Delphi.* Tradução José Carlos Barbosa dos Santos; revisão técnica Edmilson Kazwyoshi Miasaki. São Paulo: MAKRON Books do Brasil Editora Ltda., 1996. 1192p.

[Chen & Lin, 1996] **CHEN, H. & LIN, C.** *An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents.* IEEE Transactions on Systems, Man. and Cybernetics, v. 26, n. 1, February 1996, p. 1-4. **http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html**.

[Chen, 1993] **CHEN, H., LYCH, K., BASU, K & DORBIN, T.** *Generating, Integrating, and Activating Thesauri for Concept-based Document Retrieval.* IEEE Expert, abril, 1993, p. 25-34.

[Chen, 1995] **CHEN, H. et al.** *Automatic Thesaurus Generation for an Electronic Community System.* Journal of the American Society for Information Science, v. 46, n. 3, April 1995, p. 175-193. **http://ai.bpa.arizona.edu/papers/worm94/tableofcontents3_1.html**.

[Croft, 1992] **CROFT, W. Bruce & TURTLE, Howard R.** *A Comparison of Text Retrieval Models.* The Computer Journal, v. 35, n. 3, 1992, p. 279-290.

[Crouch, 1990] **CROUCH, C. J.** *An Approach to the Automatic Construction of Global Thesauri.* Information Processing and Management, v. 26, n. 5, 1990, p. 629-640.

[Ferneda, 1997] **FERNEDA, Edberto.** *Construção Automática de um Thesaurus Retangular.* Dissertação de Mestrado. Universidade Federal da Paraíba. Agosto, 1997.

[Fluhr, 1995] **FLUHR, Christian.** *Multilingual Information Retrieval.* CEA-INSTN, Saclay, France. **http://www.kgw.tu-berlin.de/~mengel/SpeechTech/ch8node7.html**.

[Frakes, 1992] **FRAKES, W.B., BAEZA-YATES, R.** *Information Retrieval – Data Structures & Algorithms.* Prentice Hall, Englewood Clifss, New Jersey, USA, 1992.

[Furtado, 1973] **FURTADO, Antonio Luiz**. *Teoria dos Grafos: Algoritmos*. Rio de Janeiro, Livros Técnicos e Científicos; São Paulo, Ed. Da Universidade de São Paulo, 1973.

[Gammoudi, 1993] **GAMMOUDI, Mohamed Mohsen.** *Méthode de Décomposition Rectangulaire d'une Relation Binaire: Une base formelle et uniforme pour la génération automatique des thésaurus et la recherche documentaire.* Thèse de Doctorat. Université de Nice – Sophia Antipolis Ecole Doctorale des Sciences pour L'Ingénieur. 1993.

[Getty, 1997] International Terminology Working Group sponsored by The Getty Information Institute. *Guidelines for Forming Language Equivalents: A model based on the Art & Architecture Thesaurus.* **Http://www.ahip.getty.edu/guidelines/index.html**.

[Godin, 1986] **GODIN, R.** *L'Utilisation de Treillis pour l'accès Aux Systèmes d'Information.* Thèse de Doctorat, Montréal, 1986.

*[Guidelines, 1997]* *Guidelines for constructing a Museum Object Name Thesaurus,* **http://www.open.gov.uk/mdocassn/holm.htm**.

[Kavanagh, 1995] **KAVANAGH, Judy.** *The Text Analyser: A Tool for Extracting Knowledge from Text.* Dissertação de Mestrado. University of Ottawa. 1995. **http://www.csi.uottawa.ca/~kavanagh**.

[Kowalski, 1997] **KOWALSKI, Gerald.** *Information Retrieval Systems: Theory and Implementation.* Klüwer Academic Publishers, 1997.

[Lancaster, 1986] **LANCASTER, F. Wilfrid.** *Vocabulary Control for Information Retrieval.* 2nd ed. (Arlington, VA: Information Resources Press, 1986), 1986, p. 218.

[Lewis, 1996] **LEWIS, D. D., JONES, K. S.** *Natural Language Processing for Information Retrieval.* Communications of the ACM, New York, v. 39, n. 1, 1996, p. 92-101.

[Lindberg, 1990] **LINDBERG, D. A . & HUMPHREYS, B. L.** *The UMLS Knowledge Sources: Tools for Building Better User Interface.* In Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November, 1990, p. 4-7.

[McGray, 1990] **McCRAY, A.T. & HOLE, W.T.** *The Scope and Structure of the First Version of the UMLS Semantic Network.* In Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November, 1990, p. 4-7.

[Monarch, 1987] **MONARCH, I. & CARBONELL, J. G.** *CoalSORT: A Knowledge-based Interface.* IEEE EXPERT, Spring, 1987, p. 39-53,

[Multites, 1997] Multisystems, *MultiTes – Thesaurus Construction*, http://www.cris.com/~multites.

[Oard & Dorr, 1996] **OARD, Douglas W. & DORR, Bonnie J.** *A Survey of Multilingual Text Retrieval.* Institute for Advanced Computer Studies and Computer Science Department. University of Maryland. UMIACS-TR-96-19 CS-TR-3615. April, 1996.

[Oard, 1997a] **OARD, Doug.** *Multilingual Information Retrieval*, http://www.ee.umd.edu/medlab/mlir/mlir_definition.html.

[Oard, 1997b] **OARD, Douglas W.** *Serving Users in Many Languages: Cross-language Information Retrieval for Digital Libraries.* D-Lib Magazine, December 1997. (http://www.dlib.org/dlib/december97/oard/12oard.html).

[Pinto e Almeida, 1995] **PINTO, Ulisses & ALMEIDA, José João Dias.** *Tratamento Automático de Termos Compostos.* Universidade do Minho, Portugal. Apresentado no XI Encontro da Associação Portuguesa de Linguística. http://www.di.uminho.pt/~jj/pln

[Pinto, 1997] **PINTO, Wilson Silva.** *Sistema de Recuperação de Informação com Navegação através de Pseudo Thesaurus.* Dissertação de Mestrado. Universidade Federal do Maranhão, 1997.

[Robredo, 1991] **ROBREDO, Jaime.** *Indexação Automática de Textos: uma Abordagem Otimizada e Simples.* Ciência da Informação, Brasília, v. 20, n. 2, Julho/Dezembro 1991, p. 130-136.

[Robredo, 1998] **ROBREDO, Jaime & CUNHA, Murilo Bastos.** *Aplicação de Técnicas Infométricas para Identificar a Abrangência do Léxico Básico que Caracteriza os Processos de Indexação e Recuperação da Informação.* Ciência da Informação, Brasília, v. 27, n. 1, Janeiro/Abril 1998, p. 11-27.

[Salton, 1972] **SALTON, Gerard.** *Experiments in Multilingual Information Retrieval.* Computer Science Dept., Cornell University. Relatório técnico 72-154, 1972. http://cs-tr.cs.cornell.edu:80/Dienst/

[Salton, 1989] **SALTON, Gerard.** *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company. 1989.

[Sosoaga, 1991] **SOSOAGA, Carmén López de.** *Multilingual Access to Documentary Database.* In A. Lichnerowicz, Editor. Proceedings of a Conference on Intelligent Text and Image Handling (RIAO91), Amsterdam, April 1991, p. 774-778.

[Yuan, 1997] **YUAN, Qunming, CHANG, Ifay.** *IT Thesaurus Construction – The Methodology and Observations.* Polytechnic Research Institute for Development and Enterprise (PRIDE), Polytechnic University. 1997. **(http://pride-i2.poly.edu/~qmyz/papers/iasted/ias-pap.html)**.

# APÊNDICE A – EXEMPLOS DE CLASSES CONCEITUAIS - DICIONÁRIO PRINCIPAL

CONCEITO = 1
DEFINIÇÃO (Contexto) = Comunicação ou notícia trazida ao conhecimento de uma pessoa ou público.
CATEGORIA = Substantivo

| Português | Francês | Italiano | Inglês |
|---|---|---|---|
| COMUNICAÇÃO | COMMUNICATION | COMUNICAZIONE | COMMUNICATION |
| ESCLARECIMENTO | INFORMATION | INFORMAZIONE | INFORMATION |
| INFORMAÇÃO | NOUVELLES | NOTIZIA | NEWS |
| INFORME | RAPPORT | | REPORT |
| NOTÍCIA | | | |

CONCEITO = 236
DEFINIÇÃO (Contexto) =Coleção de fatos ou de outros dados fornecidos à máquina a fim de se objetivar um processamento [PROC. DE DADOS].
CATEGORIA = Substantivo

| Português | Francês | Italiano | Inglês |
|---|---|---|---|
| INFORMAÇÃO | INFORMATION | INFORMAZIONE | INFORMATION |
| DADO | DONNÉE | DADO | DATA |

CONCEITO = 27
DEFINIÇÃO (Contexto) = Livro de sinônimos
CATEGORIA = Substantivo

| Português | Francês | Italiano | Inglês |
|---|---|---|---|
| DICIONÁRIO | DICTIONNAIRE | ENCICLOPEDIA | DICTIONARY |
| LÉXICO | LEXIQUE | LESSICO | LEXICON |
| THESAURUS | THESAURUS | THESAURUS | THESAURUS |
| VOCABULÁRIO | VOCABULAIRE | VOCABOLARIO | VOCABULARY |

# APÊNDICE B – ESTRUTURA DE DADOS PARA GRAFO RETANGULAR