Edberto Ferneda

Construção Automática de um Thesaurus Retangular

Campina Grande - PB
Agosto de 1997

Edberto Ferneda

Construção Automática de um Thesaurus Retangular

Dissertação apresentada ao Curso de Mestrado em Informática da Universidade Federal da Paraíba, como exigência parcial para a obtenção do grau de Mestre.

Área de concentração: Ciência da Computação

Ulrich Schiel
Orientador

Mohamed M. Gammoudi
Co-orientador

Campina Grande
1997

F364c Ferneda, Edberto
Construcao automatica de um thesaurus retangular / Edberto Ferneda. - Campina Grande, 1997.
74 f. : il.

Dissertacao (Mestrado em Informatica) - Universidade Federal da Paraiba, Centro de Ciencias e Tecnologia.

1. Banco de dados 2. Thesaurus - 3. Dissertacao I. Schiel, Ulrich, Dr. II. Gammoudi, Mohamed M., Dr. III. Universidade Federal da Paraiba - Campina Grande (PB)

CDU 004.65(043)

CONSTRUÇÃO AUTOMÁTICA DE UM THESAURUS RETANGULAR

EDBERTO FERNEDA

DISSERTAÇÃO APROVADA EM 29.08.1997

PROF. ULRICH SCHIEL, Ph.D
Presidente

PROF. MOHAMMED MOHSEN GAMMOUDI, Dr.
Examinador

PROF. CAVAN McCARTHY, Dr.
Examinador

PROF. JORGE CÉSAR ABRANTES DE FIGUEIREDO, Dr.
Examinador

CAMPINA GRANDE - PB

A meu pai, Élcio Ferneda

Sua curiosidade em conhecer novas paisagens o fez caminhar para além desta vida.

À minha mãe, Elza Ferneda

À minha esposa, Rita de Cassia

Aos meus irmãos, Edilson e Edmir

Agradecimentos

Agradeço a minha esposa, Rita de Cassia, que abdicou de alguns de seus sonhos em favor deste meu sonho que aqui se concretiza.

Agradeço minha mãe e meu pai, alicerce e apoio na construção da minha vida.

Ao professor Edilson Ferneda que, como um irmão, me incentivou, aconselhou e inspirou.

Aos professores Ulrich Schiel e Mohamed M. Gammoudi que com paciência e dedicação me orientaram durante a elaboração deste trabalho.

Ao professor Hattori e à professora Joseluce, que muito me ajudaram em momentos de dificuldade.

Agradeço também aos amigos e colegas que direta ou indiretamente me ajudaram a finalizar este trabalho: Álvaro e Ismênia, Andrea (ETFAL), Dilvan Vitor, Fabrísia e Evandro, Fernanda e Dalmer, Ianna, Gilson, Michelle, Ricardo, Sônia, Rômulo e Stênio (ETFAL), Vera Lúcia Prudência.

Ao meu irmão Edilson Ferneda que, como um professor, constantemente me ensina lições de vida.

Resumo

Pesquisas na área de Banco de Dados Documentais (BDD) mostram que o *thesaurus* é uma ferramenta bastante útil para a indexação e recuperação de informações textuais. Na maioria dos BDDs a construção do *thesaurus* é feita de forma manual. Neste trabalho é apresentado um sistema para a construção automática de *thesaurus*.

Este trabalho é baseado em dois outros trabalhos de pesquisa. O primeiro diz respeito à primeira etapa da construção de um *thesaurus*: a extração de termos de um conjunto de documentos. O segundo refere-se ao processo de organização dos termos e dos documentos utilizando um modelo conceitual. A unificação destes dois trabalhos permitiu a implementação de um sistema para a construção automática de *thesaurus*.

Abstract

The researches in the field of Textual Databases indicate that the *thesaurus* is a very useful tool for indexing and retrieving textual information. In the great majority of Textual Databases, the *thesaurus* construction is not done in a automatic way. This work presents a system for the automatic construction of a *thesaurus*.

This work is based on two other researches. The first research is concerned with the initial steps for building a *thesaurus*: the extraction of terms from a set of documents. The second research deals with the process of organizing the terms and the document using a conceptual model. The unification of these two researches led to the implementation of a system used for the automatic construction of *thesaurus*.

Sumário

Figuras

1. Introdução

Quando Johann Gutenberg inventou o tipo móvel e apresentou a primeira prensa na Europa, mudou definitivamente a cultura ocidental. Ele levou dois anos para compor os tipos de sua primeira Bíblia, mas feito isso teve condições de imprimir centenas de exemplares.

O invento de Gutenberg permitiu não apenas a multiplicação de cópias, possibilitou também a criação de uma interface padronizada: Sumários, capítulos, cabeçalhos, numeração de páginas, índices, referências; todos esses dispositivos classificatórios sustentando-se uns aos outros no interior de uma estrutura sistêmica. Com tais recursos somos capazes de um exame rápido do conteúdo de uma obra, de um acesso seletivo e não linear ao texto, de conexões a uma infinidade de outros textos.

Porém, quando estamos diante de um computador, a pesquisa direta a informações torna-se um trabalho extremamente árduo. Essas dificuldades são parcialmente compensadas por algumas características de interface que se disseminaram durante os anos oitenta: a representação através de ícones, o uso do *mouse*, os menus, os recursos gráficos, o hipertexto e a hipermídia. [Lévy 1993; Martin 1992].

Apesar do rápido aumento da velocidade e capacidade dos computadores, cada vez mais notamos o quanto somos incapazes de manipular de maneira satisfatória a quantidade de informações a nossa disposição. Novos relatórios, livros, revistas, filmes, e todas as novas mídias continuam gerando informações em uma quantidade sempre crescente. Os recursos

para o armazenamento e recuperação de informações que possuímos hoje não são comparáveis às facilidades de produção e proliferação de informações.

As informações podem ser divididas em dois grandes grupos: *informações bem estruturadas* e *informações mal estruturadas*. O sucesso dos sistemas gerenciadores de bancos de dados deve-se em grande parte pelo uso de informações bem estruturadas. Mas informações precisamente pré-definidas e quantificadas, tal como nos sistemas de banco de dados, pertencem a um universo específico. Muitos dos campos de interesse dos sistemas de informação se caracterizam por uma mistura de informações mal estruturadas e dados bem estruturados.

Particularmente, as informações sob forma de texto são de fundamental importância para a sociedade, antes mesmo do invento de Gutenberg. E é no domínio das informações textuais que se concentra este trabalho.

Para uma recuperação eficaz, as informações disponíveis devem estar bem descritas. A recuperação precisa das informações depende da análise correta e consistente do conteúdo do material a ser pesquisado e recuperado. Uma das formas de prover esta análise é por meio de um mecanismo de indexação. O objetivo da indexação é 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.

O processo de indexação consiste em extrair termos significativos de documentos, e organizá-los utilizando um modelo conceitual. Uma das formas de indexação mais conhecidas e utilizadas nos sistemas de recuperação de informações é o *thesaurus*.

O termo *thesaurus* teve origem no ano de 1852 com a publicação do dicionário analógico de Peter Mark Roget, entitulado "*Thesaurus of English words and phrases*". Nesse dicionário as palavras não foram agrupadas em ordem alfabética, como nos dicionários tradicionais, mas de acordo com as idéias que elas exprimem. Segundo Roget, o propósito de um dicionário comum é simplesmente explicar o significado das palavras. O que ele pretendia era, tendo-se a idéia, encontrar as palavras pelas quais essa idéia pudesse ser expressa de maneira mais adequada.

A aplicação do *thesaurus* na recuperação de informações surgiu da necessidade de manipular grandes quantidades de documentos. Os documentalistas se utilizam de um

thesaurus para indexar manualmente os documentos ou livros de forma a encontrar rapidamente uma obra que diz respeito a um determinado assunto. Isso dá uma outra forma de acesso às obras além do acesso através de seus autores ou títulos. Uma simples classificação das obras por temas em uma biblioteca é um primeiro resultado da utilização de *thesaurus*. A seguir apareceram os arquivos manuais que permitiam recuperar obras indexadas por palavras-chave. O *thesaurus* era então utilizado como elemento de normalização, fornecendo uma lista dos termos a serem considerados para indexar um documento. Estruturas hierárquicas de *thesaurus* surgiram rapidamente: no momento da classificação da obra por tema, introduziu-se os subtemas. Por exemplo, sendo o tema *matemática* muito genérico, este foi decomposto em subtemas: *Estatística*, *probabilidade*, *teoria dos conjuntos*, etc.

Um *thesaurus* é constituído de um conjunto de termos em linguagem natural e um conjunto de relações semânticas hierárquicas entre esses termos, formando uma rede. Sua riqueza é maior ou menor de acordo com a escolha dos termos (palavras chave ou grupos mais complexos), a natureza e o número de relações semânticas. Com esta estrutura os usuários podem refinar suas consultas e obter respostas relevantes. A integração de um *thesaurus* em um Sistemas de Recuperação de Informações (SRI) é fundamental para o aumento de sua eficiência.

Atualmente o *thesaurus* é usado na maioria dos SRIs. Eles podem ser construídos de forma manual ou automaticamente através de técnicas computacionais.

A construção manual de *thesaurus* possui duas principais desvantagens:

- Requer muito tempo e vários especialistas, o que torna esse processo extremamente custoso.
- A subjetividade é um grande empecilho. A partir de um mesmo documento, dois especialistas podem extrair diferentes conjuntos de termos que consideram relevantes para a descrição dos assuntos contidos no texto. E mesmo partindo-se de um mesmo conjunto de termos, dois especialistas podem construir diferentes thesaurus.

O objetivo deste trabalho é apresentar um método para a construção automática de *thesaurus*. O método permite a construção de um *thesaurus* sem conhecimento prévio do conjunto de documentos tratado e permite a construção automática de relações semânticas que

ultrapassam, por sua riqueza, as relações hierárquicas clássicas da maioria dos *thesaurus* manuais

O primeiro passo para a construção de um *thesaurus* em computador consiste na extração de termos a partir de um conjunto de textos. No Capítulo 2 é descrito um método para a extração automática de termos. O método é baseado nos trabalhos de pesquisa realizados no *Laboratoire de Génie Informatique de Grenoble* - França, durante a década de 80 [Bruandet 1989b].

Após a extração dos termos, a construção do *thesaurus* é feita de forma automática, utilizando o método de ***Decomposição Retangular de uma Relação Binária*** [Gammoudi 1993]. No Capítulo 3 é apresentada a fundamentação matemática utilizada na construção de um ***Thesaurus Retangular*** e na pesquisa a uma base de dados documental.

Os termos extraídos dos textos podem ser representados como uma relação binária entre dois conjuntos de termos. Aplicando-se o método de Decomposição Retangular é obtido como resultado um conjunto de *retângulos ótimos*. Um *Thesaurus* Retangular é um grafo de retângulos ótimos gerado através da aplicação de uma relação de ordem sobre o conjunto de retângulos ótimos. O Capítulo 4 descreve os passos para a construção automática do *Thesaurus* Retangular.

A representação da base de documentos é feita através de uma relação binária cujo domínio é o conjunto dos descritores dos documentos (termos) e o codomínio é o conjunto de referências aos documentos. Para classificar os documentos e seus descritores é utilizado o mesmo método de Decomposição Retangular utilizada na construção do *thesaurus*. Aplicando-se também a mesma relação de ordem utilizada na construção do *thesaurus*, os retângulos ótimos obtidos são organizados na forma de um grafo. No Capítulo 5 são apresentadas as etapas envolvidas na organização hierárquica dos documentos.

No Capítulo 6 são propostas diferentes operações de pesquisa que podem ser realizadas em um ***Sistema Documental Retangular***, de acordo com a maneira pela qual a consulta é expressa. Como ferramenta matemática de auxilio às consultas é utilizada a noção de ***Conexão de Galois***.

Os trabalhos de Marie-Françoise Bruandet [Bruandet 1989b] e o trabalho de Mohamed Gammoudi [Gammoudi 1993], fornecem uma sólida base teórica para a implementação de um

sistema documental. Aplicando-se a maioria das idéias apresentadas nesses dois trabalhos, foi desenvolvido um sistema utilizando linguagem **Delphi 3.0** [Pacheco, 1996; Henderson, 1996]. No Capítulo 7 é apresentado o sistema assim como alguns resultados obtidos em experimentações realizadas.

Alguns novos trabalhos, direta ou indiretamente ligados a este, podem ser propostos para o aperfeiçoamento da ferramenta que foi desenvolvida. No Capítulo 8 são apresentadas as conclusões deste trabalho e são sugeridos alguns assuntos que poderão ser abordados em futuros projetos.

2. Um Método para Extração de Termos

Neste capítulo é apresentado um método para a extração automática de termos a partir de textos em linguagem natural. O método é baseado em trabalhos de Marie-Françoise Bruandet do *Laboratoire Génie Informatique de Grenoble*. Os estudos de Bruandet foram publicados em uma série de artigos ao longo de quase uma década [Bruandet 1980a, 1980b, 1981, 1982a, 1982b, 1985, 1989a, 1989b] e baseiam-se em resultados de pesquisas apresentados em [Attar 1977].

O método é baseado na pesquisa de associações entre pares de palavras. Por exemplo, considere um texto onde aparecem as três frases seguintes:

```
... o banco de dados...
      └──────┘
        d=2

..os bancos de dados relacionais são...
     └──────┘
       d=2

...as consultas em um banco de dados feitas através...
                    └──────┘
                      d=2
```

Em cada uma das frases a distância (número de palavras) entre o termo "banco" e o termo "dados" é igual a 2. Isso que significa que existe no texto uma forte ligação semântica entre esses dois termos, ou pelo menos eles estão relacionados a uma mesma noção.

Para cada par de palavras é calculado um valor que expressa a força de ligação entre elas. Para isso considera-se:

- a proximidade contextual (a distância **d**) entre cada par de palavras que aparecem uma mesma frase;
- a categoria gramatical de cada uma das palavras;
- a freqüência com que essas duas palavra aparecem juntas.

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 será considerado 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. Os valores dessas medidas são armazenadas em uma matriz termo-termo. A partir do Grafo representado por essa matriz são extraídos os *cliques* (subgrafos máximos completos), que representam as idéias contidas nos textos.

2.1 Medida Utilizada nas Ligações Termo-Termo

Seja **T** o conjunto de termos extraídos dos documentos. O primeiro elemento a ser definido é uma medida para avaliar a ligação contextual entre dois termos. A i-ésima ocorrência de um termo ***x*** do vocabulário **T**, simbolizado por $\mathbf{w_x(i)}$, é definido por suas coordenadas:

$$w_x(i) = \langle ND_x(i), NF_x(i), NT_x(i) \rangle \quad (1)$$

onde $ND_x(i)$ *designa o número do documento;* $NF_x(i)$ *o número da frase no documento;* $NT_x(i)$ *a posição do termo na frase.*

Para cada par de termos (x, y), define-se uma distância **d** entre a i-ésima ocorrência de ***x*** e a j-ésima ocorrência de ***y***:

(2)

$$d(w_x(i), w_y(j)) = \begin{cases} |NT_x(i) - NT_y(j)| & se \begin{cases} ND_x(i) = ND_y(i) \\ e \\ NF_x(i) = NF_y(i) \end{cases} \\ 0, \textit{caso contrário} \end{cases}$$

Seja **F** uma função que expressa a força de ligação entre dois termos. Ela é definida como sendo o inverso da distância **d** definida acima:

(3)

$$F(w_x(i), w_y(j)) = \begin{cases} \dfrac{1}{d(w_x(i), w_y(j))} & se\ d(w_x(i), w_y(j)) \leq t \\ 0, \textit{caso contrário} \end{cases}$$

t *é um limite fixado experimentalmente, e será descrito a seguir.*

As associações entre certas categorias de palavras (por exemplo, entre duas preposições) não possuem interesse. Na equação 3, acima, foi introduzido um limite ***t*** para a distância entre dois termos. Esse limite considera a categoria gramatical dos termos. Um adjetivo e um substantivo, por exemplo, são considerados como relacionados semanticamente somente se aparecerem a uma distância igual ou inferior à um determinado valor.

O limite ***t*** pode ser formalizado da seguinte maneira:

(4)

$$t(x, y) = \text{LIMITE}[\ \text{CAT}(x), \text{CAT}(y)\ ]$$

onde LIMITE *é a distância máxima entre duas categorias gramaticais e* CAT *é uma função que retorna a categoria gramatical de um termo.*

Os valores de ***t*** são definidos para as categorias gramaticais consideradas em uma determinada aplicação e podem ser representadas através de um grafo. O grafo da Figura 2.1 apresenta um exemplo onde estão representadas algumas categorias gramaticais (*substantivos*, *adjetivos*, *verbos* e *preposições*), e o limite ***t*** entre elas.

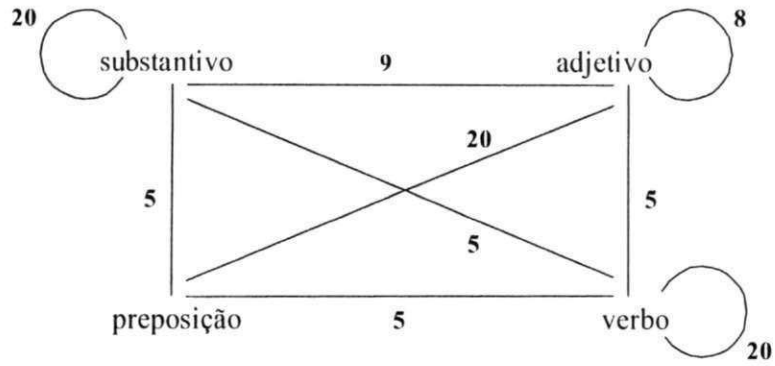


Figura 2.1: Grafo das distâncias entre categorias gramaticais

As categorias gramaticais consideradas e a distância máxima entre elas podem ser modificadas em função de interesses específicos da aplicação ou dos resultados obtidos. O limite ***t*** intervém no cálculo da função **F**. Uma definição mais completa da função **F** é dada por.

(5)

$$F(w_x(i), w_y(j)) = \begin{cases} \dfrac{1}{d(w_x(i), w_y(j))} & \textit{se } d(w_x(i), w_y(j)) \le t(x, y) \\ 0, \textit{caso contrario} \end{cases}$$

com $t(x, y) = \text{LIMITE}[\, CAT(x), CAT(y) \,]$

Seja **b** a função do somatório de **F** para todas as ocorrências ***i*** e ***j*** de ***x*** e ***y***:

(6)

$$b(x, y) = \sum_i \sum_j F(w_x(i), w_y(j))$$

Uma definição preliminar da medida de associação entre ***x*** e ***y*** é dada por:

(7)

$$M_1(x, y) = \frac{b(x, y)}{f(x, y)}$$

onde $f(x, y)$ *é o número de ocorrências (freqüência) do par* (x, y) *no conjunto de documentos.*

$$0 \le M_1(x, y) \le 1$$

Utilizando apenas $M_1(x, y)$, a força de ligação poderá valer 1 quando dois termos *x* e *y* forem adjacentes em uma mesma frase e só aparecerem uma única vez no conjunto de documentos. Obviamente esses pares de termos não devem ser considerados, representando uma exceção significativa na utilização da medida M_1. Para eliminar tal problema é

introduzido um fator de correção **k** (**k|0, 1|**), que é função da freqüência f(x, y) do par (x, y). O fator **k** recebe o valor zero quando f(x, y) vale 1 e tende a 1 conforme f(x, y) aumenta. Assim, uma nova medida M_2 é definida como:

$$M_2(x, y) = k(x, y) \times M_1(x, y) \qquad (8)$$

$$com \quad k(x, y) = \frac{(f(x, y) - 1)^n}{f(x, y)^n}$$

onde **n** *é um parâmetro inteiro definido por experimentação.*

Um aumento de ***n*** tende a reforçar as ligações muito freqüentes e atenuar as ligações pouco freqüentes. O parâmetro ***n*** tem seu valor definido de forma empírica, podendo ser utilizado para ajustar a aplicação de acordo com finalidades específicas. Nos experimentos apresentados em [Bruandet 1989b] foi utilizado o valor ***n*** = 2.

A medida $M_2(x, y)$ pode ser também representada através da seguinte fórmula geral:

$$M_2(x, y) = \frac{\sum_i \sum_j F(w_x(i), w_y(j))}{f(x, y)} \times \frac{(f(x, y) - 1)^n}{f(x, y)^n} \qquad (9)$$

A introdução de diferentes parâmetros permite uma melhor adaptação da medida M_2 às necessidades da aplicação. Através do fator corretivo ***k*** é possível atuar sobre a freqüência com a qual os pares de termos devem se relacionar. O limite ***t*** permite selecionar os termos cuja categoria gramatical é interessante para qualificar o contexto de um termo. Além da melhoria qualitativa, estes parâmetros permitem reduzir consideravelmente o número de termos selecionados e a quantidade de ligações entre eles.

2.2 Construção da Matriz Termo-Termo

Os valores da medida M_2 são armazenadas em uma **matriz termo-termo**, como mostrado no exemplo da Figura 2.2.

Termo	p	t	u	w	x	y	z
p	-	0,12	0	0	0	0,18	0
t	-	-	0,21	0,19	0,23	0,18	0,17
u	-	-	-	0,18	0,13	0	0
w	-	-	-	-	0	0	0
x	-	-	-	-	-	0,27	0,26
y	-	-	-	-	-	-	0,23
z	-	-	-	-	-	-	-

Figura 2.2: Matriz termo-termo contendo os valore da medida M_2

É importante observar a simetria da matriz gerada, além do fato da diagonal ser desconsiderada. Essas características permitem uma redução significativa no espaço de armazenamento das informações contidas na matriz.

A partir da matriz contendo os valores de M_2 (Figura 2.2) constrói-se a **matriz binária termo-termo**. Neste processo é possível eliminar as ligações mais fracas, de acordo com um limite mínimo preestabelecido. A matriz binária apresentada na Figura 2.3, por exemplo, é construída a partir da matriz da Figura 2.2, utilizando um limite mínimo igual a 0,14. Dessa forma são eliminadas as ligações mais fracas { (t, p), (x, u) }, melhorando a "qualidade" das ligações entre os termos e reduzindo ainda mais a quantidade de informações a serem armazenadas.

Termo	p	t	u	w	x	y	z
p	-	0	0	0	0	1	0
t	-	-	1	1	1	1	1
u	-	-	-	1	0	0	0
w	-	-	-	-	0	0	0
x	-	-	-	-	-	1	1
y	-	-	-	-	-	-	1
z	-	-	-	-	-	-	-

Figura 2.3: Matriz binária termo-termo

2.3 Cliques

Considerando o grafo associado à matriz binária termo-termo (Figura 2.4), constata-se que é difícil interpretar e analisar as informações nela contidas. A solução para isto está em selecionar, a partir deste grafo, os subgrafos máximos completos, chamados **cliques** (subgrafos cujos nós estão todos conectados entre si) Figura 2.5.

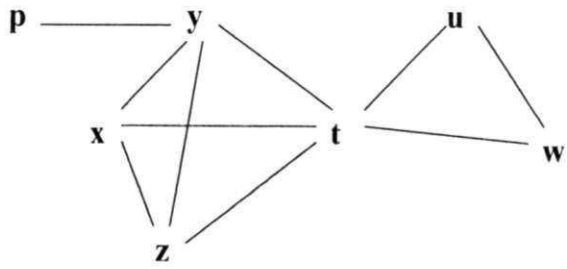


Figura 2.4: Grafo representando a matriz binária termo-termo

Diversos algoritmos para a extração de cliques de um grafo estão disponíveis no domínio da teoria dos grafos. Um deles é apresentado em [Reingold 1977] (Anexo A).

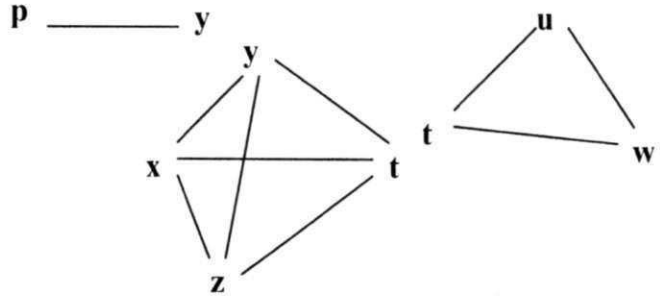


Figura 2.5: Cliques extraídos da matriz binária termo-termo

Através dos cliques obtém-se uma representação da matriz termo-termo sem perda de informações. Cada clique fornece um conjunto de termos completamente conectados entre si. Pode-se dizer que cada clique representa um conceito ou uma idéia contida no conjunto de textos.

2.4 Experimentação

Em [Bruandet 1989b] são apresentados alguns resultados obtidos em uma experimentação realizada com um texto contendo 15 capítulos, com um total de 70350 palavras.

Os resultados obtidos foram validados através da comparação com a indexação manual efetuado por documentalistas. 80% dos termos selecionados manualmente foram também encontrados pela indexação automática.

A análise dos cliques mostra que as informações representativas são essencialmente veiculadas pelos substantivos e por certos adjetivos. Os verbos são muito pobres do ponto de vista semântico, pois tratam-se essencialmente de verbos auxiliares. As preposições são relevantes apenas para certos tipos de aplicação.

3. Thesaurus Retangular: Fundamentação Matemática

Para a construção do *thesaurus* será utilizado o método de ***Decomposição Retangular de uma Relação Binária***, apresentado em [Gammoudi 1993]. Neste trabalho estão os formalismos matemáticos e os algoritmos necessários para a geração automática de um *Thesaurus* "Retangular".

O principal formalismo utilizado é o das ***Relações Binárias***. Uma relação binária pode ser decomposta em um conjunto mínimo de *retângulos ótimos*. Essa decomposição mostrou-se pertinente em vários domínios da informática. Um exemplo é a estruturação de uma base de dados documental [Belkhiter 1992].

Um outro formalismo é o ***Reticulado de Galois***. Após ter sido utilizado em teoria das ordens [Mac Neille 1937] e em automatismo para a resolução de problemas de ordenamento [Baptiste 1984], foi recentemente utilizado em análise combinatória de dados [Guenoche 1987] e aplicado em Inteligência Artificial [Wille 1985].

3.1 Relação Binária

Uma relação binária é uma sentença aberta (ou função proposicional) de duas variáveis, P(x, y), que pode ser verdadeira ou falsa para cada par (x, y) de elementos de um universo do discurso. Por exemplo:

x* está relacionado com *y

Em terminologia gramatical vemos que a sentença aberta acima é constituída de um "molde" de uma sentença ("... está relacionado com ...") e de duas variáveis (x e y), que resulta em uma sentença que podemos classificar como verdadeira ou falsa [Abe 1992].

Definição 3.1

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 × F.

$$R = \{ (x, y) \in E \times F \mid xRy \}$$

Indica-se por **xRy** *o fato de um elemento* **x** *de* E *estar ligado a um elemento* **y** *de* F *através da relação* **R**.

Um elemento de uma relação R é um par ordenado (x, y). Diz-se que x é um argumento (ou entrada) de **R** e que y é uma imagem (ou saída) de x através de **R**.

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

Para uma relação R pode-se associar os seguintes conjuntos:

- Conjunto imagem de x: $x.R = \{y \mid xRy\}$
- Conjunto dos antecedentes de y: $R.y = \{ x \mid xRy \}$
- Domínio de R: $dom(R) = \{ x \mid \exists y: xRy \}$
- Codomínio de R: $cod(R) = \{ y \mid \exists x: xRy \}$

Sendo R uma relação binário definida sobre os conjuntos E e F, pode-se definir as seguintes operações

- Inverso : $R^{-1} = \{ (y, x) \in F \times E \mid yRx \}$
- Composição: $R \circ R' = \{(x, y) \in E \times F \mid \exists t: xRt \ \& \ tR'y\}$
- Interseção: $R \cap R' = \{ (x, y) \in E \times F \mid xRy \ \& \ xR'y \}$
- União: $R \cup R' = \{ (x, y) \in E \times F \mid xRy \vee xR'y \}$

  onde "$\vee$" é o símbolo de adição lógica.

Propriedade 3.1

Quanto mais imagens uma relação associar a uma entrada, menos determinista ela será. De maneira formal, se R *e* R' *são duas relações binárias sobre um conjunto* E, *diz-se que* R *é* **mais determinista** *que* R' *se e somente se* $R^{-1} \circ R \subseteq R'^{-1} \circ R'$, *onde o símbolo "$\circ$" representa o operador de composição das relações.*

Propriedade 3.2

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

3.2 Retângulo de uma Relação Binária

Em uma terminologia proveniente da teoria dos grafos, uma relação binária R de E em F, define os arcos de um *grafo bipartido* sobre $E \cup F$. A partir daí, um retângulo é um *subgrafo bipartido completo* (ou *clique*) do grafo $(E \cup F, R)$, enquanto um *retângulo máximo* é um *subgrafo bipartido completo máximo* (ou *clique máximo*).

No domínio da análise combinatória de dados, o retângulo é conhecido com o nome de *sub-matriz completa* enquanto o retângulo máximo é chamado *sub-matriz completa primeira*.

Definição 3.2

Seja R *uma relação binária definida de* E *em* F. *Um* **retângulo** *de* R *é um par 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 $R^{++} = \text{dom}(R) \times \text{cod}(R)$.

3.2.1 Retângulo Máximo

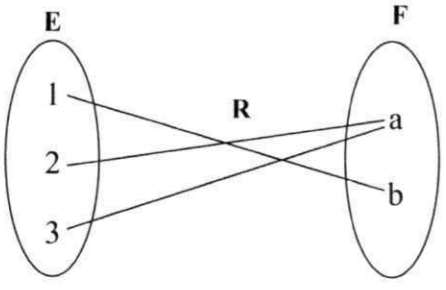
Definição 3.3

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'):

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

Exemplo

Seja a relação binária R *definida de* E *em* F *como na figura abaixo.*



O conjunto C *de retângulos da relação R é:*

$$C = \{ \mathbf{(\{1\}, \{b\})}, (\{2\}, \{a\}), (\{3\}, \{a\}), \mathbf{(\{2, 3\}, \{a\})} \}$$

Utilizando a definição de Retângulo Máximo (Definição 3.3), obtém-se o seguinte conjunto C_{max} *de retângulos máximos:*

$$C_{max} = \{ (\{1\}, \{b\}), (\{2, 3\}, \{a\}) \}$$

Através de uma representação matricial é fácil visualizar e identificar os retângulos máximos da relação binária R:

	a	b
1	0	1
2	1	0
3	1	0

Proposição 3.1

Seja R *uma relação binária finita e* (a, b) ∈ R. *A união dos retângulos de* R *que contém o elemento* (a, b) *é igual à relação (Figura 3.1):*

$$\Phi_R(a, b) = \mathbf{I}(b.R^{-1}) \circ R \circ \mathbf{I}(a.R).$$

3.2.2 Relação Elementar

Dizemos que a relação $\Phi_R(x, y)$, tal como definida acima, é uma ***relação elementar*** contendo o elemento (x, y) de R. A título de ilustração, a Figura 3.1-d é a relação elementar contendo o elemento (a, 1) da relação inicial R ilustrada pela Figura 3.1-a.

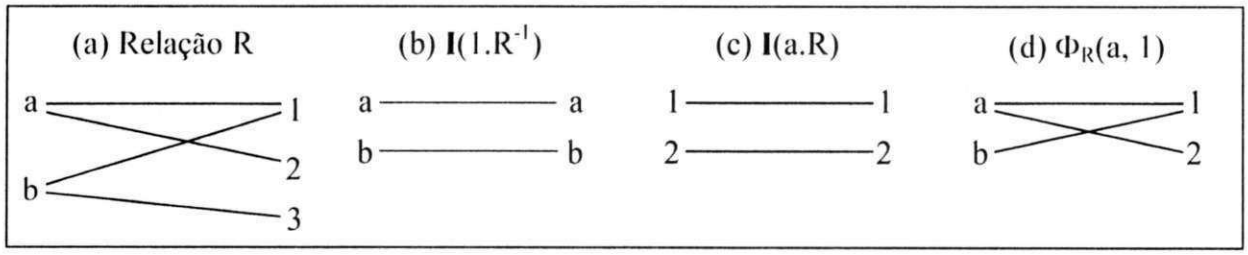


Figura 3.1: A noção de Relação Elementar

Utiliza-se a notação "PR" para designar uma relação elementar, e "RE" para designar um retângulo. Em particular, RE(a, b) designa um retângulo contendo o elemento (a, b).

3.2.3 Ganho de Espaço de Armazenamento

Se i = cardinal(A) e j = cardinal(B), torna-se possível, para cada retângulo RE = (A, B) de R, substituir i × j pares (Figura 3.2a) por i + j pares (Figura 3.2b). A partir daí medimos o ganho de espaço de codificação de informação por $g_{i,j} = ((i \times j) - (i + j))$

Desde que $g_{i,j} \geq 0$, o que é atingido para i > 1 e j > 1, obtemos um ganho apreciável em espaço de representação da informação.

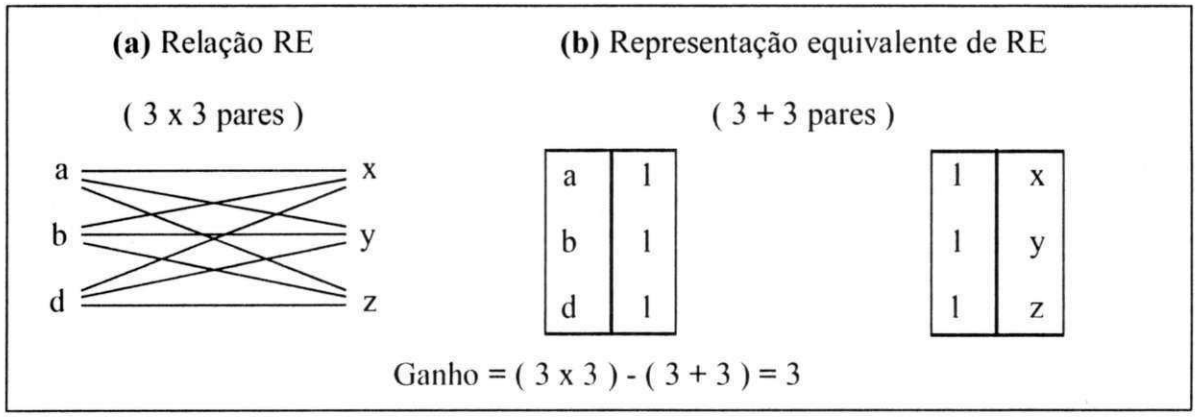


Figura 3.2: Representações equivalentes de um mesmo retângulo

Para i > 2 e J > 2, $g_{i,j} > 0$ e cresce em função de i e j. No entanto, o desperdício (ganho negativo) não pode jamais ultrapassar o valor 1, atingido para i = 1 ou j = 1. Pode-se representar todos os elementos do retângulo RE pelos pares (a, c) e (c, y), onde **c** é uma

constante que identifica o retângulo RE (c = 1 na Figura 3.2), **a** representa um elemento qualquer do conjunto A, e **y** representa um elemento qualquer de B. Na Figura 3.2, que ilustra os dois modos de representação possíveis de um mesmo retângulo, temos RE = ({a, b, d}, {x, y, z}).

3.2.4 Retângulo Ótimo

Definição 3.4

Um retângulo contendo um elemento (a, b) *de uma relação* R *é dito* **ótimo** *se ele produz o máximo de ganho entre todos os retângulos máximos que contém* (a, b) *(Figura 3.3).*

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

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

A Figura 3.3a apresenta um exemplo de uma relação R. As Figuras 3.3b, 3.3c e 3.3d representam os três retângulos máximos de R que contém o elemento (y, 3). Os ganhos de espaço obtidos com esses três retângulos máximos são respectivamente 1, 0 e -1. Portanto, o retângulo ótimo que contém o elemento (y, 3) de R é o retângulo ilustrado pela Figura 3.3b, uma vez que ele fornece um ganho máximo.

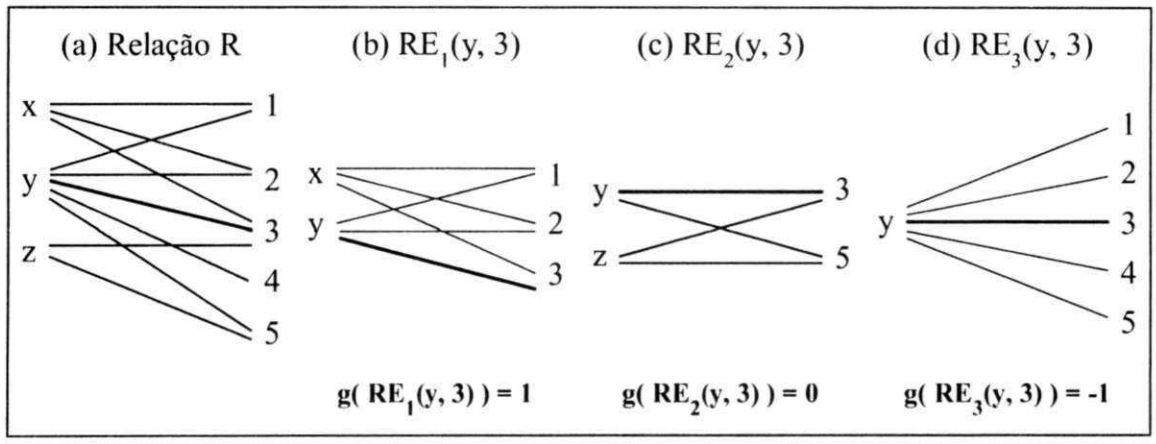


Figura 3.3: Retângulo ótimo

No Anexo B é apresentado o algoritmo de extração de retângulos ótimos de uma relação binária.

3.2.5 Cobertura de uma Relação

Definição 3.5

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

3.2.6 Cobertura Mínima de uma Relação

Definição 3.6

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

Em [Gammoudi 1993] é proposto um método heurístico que permite a obtenção da cobertura mínima de uma relação binária. O algoritmo é apresentado no Anexo C.

3.3 Relação Difuncional

O não-determinismo de uma relação induz a uma certa desordem na associação das imagens aos argumentos. Essa desordem é devido ao fato que dois argumentos quaisquer podem compartilhar certas imagens e não compartilhar outras (Figura 3.4b). As relações difuncionais são relações que guardam uma uniformidade na associação das imagens aos argumentos (Figura 3.4a).

Definição 3.7

Uma relação binária R *de* E *em* F *é dita* **difuncional** *(ou regular) se e somente se:*

$$R \circ R^{-1} \circ R = R.$$

Em [Everett 1944] é proposta a seguinte definição de uma relação difuncional.

Definição 3.8

Uma relação R *de* E *em* F *é dita* **difuncional** *se e somente se:*

$$\forall a, b \in E: a.R \cap b.R \neq \varnothing \Rightarrow a.R = b.R$$

Exemplo

Sejam E={w, x, y, z }, F={1, 2, 3} *e* R={(w, 1), (w, 2), (x, 1), (x, 2), (y, 3), (z, 3)}. R *é difuncional porque ela respeita a última definição, isto é,* w.R=x.R={1, 2} *e* y.R=z.R={3} *(Figura 3.4a).*

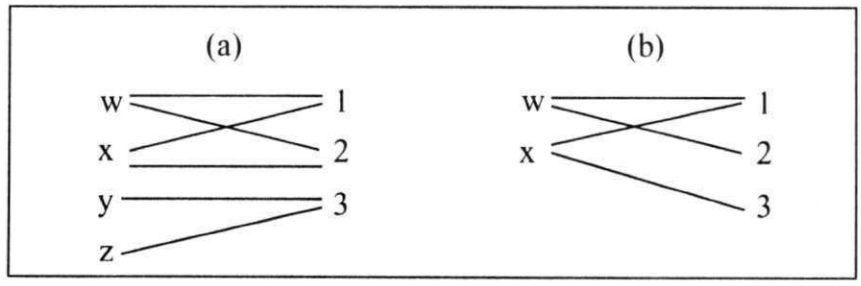


Figura 3.4: Ilustração do conceito de Relação Difuncional

Exemplo

Seja E={w, x}, F={1, 2, 3} *e* R={ (w, 1), (w, 2), (x, 1), (x, 3) }. R *não é difuncional porque* $w.R \cap x.R = \{1\}$ *mas* $w.R \neq x.R$ *(Figura 3.4b).*

Seja R a relação difuncional ilustrada pela Figura 3.5a. R pode ser decomposta segundo a união dos dois retângulos disjuntos RE_1 e RE_2, ilustrados respectivamente pelas Figuras 3.5b e 3.5c.

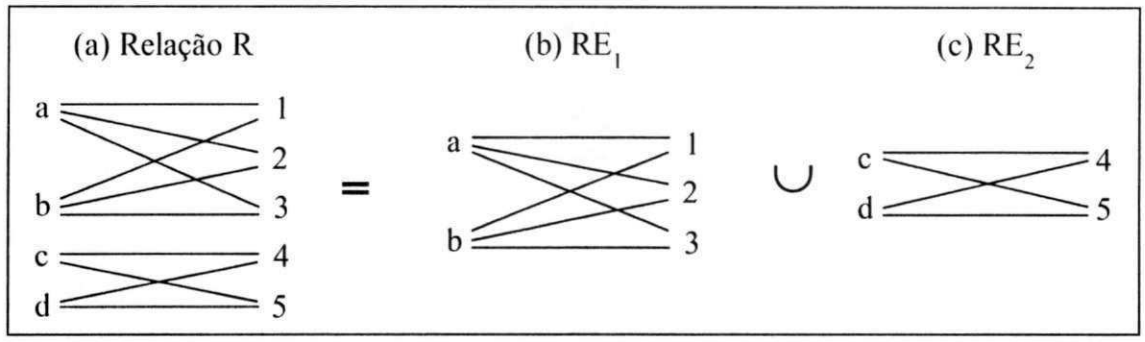


Figura 3.5: Decomposição canônica de uma relação difuncional

Em [Jaoua 1992a] é mostrado que a decomposição canônica de uma relação difuncional R em uma união de retângulos disjuntos (Figura 3.5) permite uma redução apreciável das redundâncias, assim como uma melhor organização dos dados.

3.4 Relação Retangular

Definição 3.9

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

Observações

1. Existe uma correspondência biunívoca entre os retângulos (A_i, B_i) e as relações retangulares $A_i \times B_i$, salvo quando $A_i = \varnothing$ ou $B_i = \varnothing$. Os retângulos $(\varnothing, B_1)$ e $(\varnothing, B_2)$, por exemplo, correspondem à relação retangular $\varnothing$. Esta é a razão pela qual distinguimos os retângulos das relações retangulares. Como veremos a seguir, os retângulos nos permitirão chegar a uma estrutura de reticulado.

2. Com a finalidade de não tornar o texto muito complexo, a partir de agora será utilizada a nomenclatura "*retângulo* (A, B) *contendo um elemento* (a, b) *de uma relação* R" ao invés de "*relação retangular* $A \times B$ *de* R, *associada ao retângulo* (A, B), *e contendo um elemento* (a, b) *de* R", que seria mais preciso.

3.4.1 Propriedades

Nesta seção serão apresentados diversas propriedades formais das relações retangulares. As provas destas propriedades podem ser encontradas em [Gammoudi 1993].

Proposição 3.2

O inverso de uma relação retangular é uma relação retangular

Proposição 3.3

A interseção e a composição de duas relações retangulares é uma relação retangular.

Segundo as duas proposições acima pode-se deduzir que se R_1 e R_2 são relações retangulares, então $R_1 \circ R_2^{-1}$ e $R_1^{-1} \circ R_2$ são também relações retangulares.

3.5 Reticulado de Retângulos

Nesta seção serão revistas as definições de *relação de ordem* e de *reticulado* para demonstrar que o conjunto de retângulos máximos forma um ***Reticulado***. Será apresentada

também a *Conexão de Galois*, que vai servir de ferramenta formal para reformular as consultas a uma base de documental.

3.5.1 Relação de Ordem e Reticulado

Nesta seção lembraremos a definição de **Relação de Ordem** e de **Reticulado**. Será demonstrado que o conjunto de retângulos máximos forma um **Reticulado de Galois**.

Definição 3.10

Seja R *uma relação binária definida sobre um conjunto* E. *Diz-se que* R *é uma* **relação de ordem** (parcial) *se e somente se são verificadas as seguintes condições:*

- Reflexividade: $\forall x \in E$, tem-se xRx
- Anti-simetria: $\forall x, y \in E$, se xRy e yRx, então x = y
- Transitividade: $\forall x, y, z \in E$, se xRy e yRz, então xRz

Uma relação de ordem R é geralmente indicada com o símbolo $\leq$ (leia-se "menor ou igual" ou "precede"); assim, $x \leq y$ significa que "x é menor ou igual a y" ou "x precede y".

Definição 3.11

Uma ordem é dita ***total*** *se* $\forall (x, y) \in E \times E$, *tem-se xRy ou yRx.*

Uma ordem total é também chamada *ordem completa*, *ordem linear*, ou ainda *cadeia*.

Exemplo

A relação de ordem *em* E = { 1, 2, 3 }:

$$R = \{ (1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3) \}$$

Temos 1R2, 1R3, 2R3, *isto é, dois elementos quaisquer de* E *são comparáveis pela ordem* R. *Logo,* R *é uma relação de ordem total em* E.

Exemplo

A ordem de divisibilidade no conjunto N *dos inteiros naturais positivos,* N = {1, 2, 3...} *denotado por* x | y (x «divide» y) *é reflexivo (o quociente* x|x *é igual a* 1*), anti-simétrico e transitivo. Ela é portanto parcial. A Figura 3.6 representa a organização dos divisores de 36 conforme a relação de ordem "divide". Podemos constatar que 12 não divide 18 e que 18 não divide 12. A ordem então não é total.*

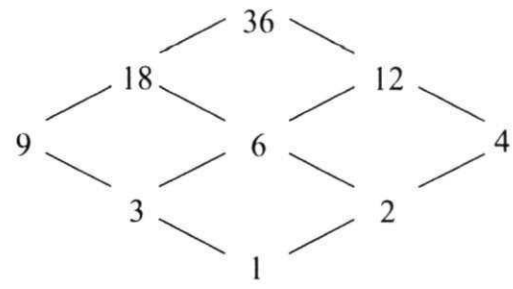


Figura 3.6: Reticulado dos divisores de 36

Seja (R, ≤) e X ⊆ R. Um limite superior de X é um elemento s ∈ R tal que para todo elemento x ∈ X, tem-se x ≤ s. O menor limite superior de X (supremo de X) é um elemento p ∈ R tal que p é um limite superior de X e para todo limite superior s de X, tem-se p ≤ s. As noções de limite inferior de X e do maior limite inferior de X (ínfimo de X) são analogamente definidas.

Definição 3.12

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

Seja (R, ≤) um reticulado, e X ⊆ R. Chama-se ▼X o supremo de X e ▲X o ínfimo de X. No caso onde X possui dois elementos, utiliza-se uma notação infixa, isto é, s▼r em lugar de ▼{s, r}. e s▲r em lugar de ▲{s, r}.

Definição 3.13

Um reticulado (R, ≤) *é dito* **distributivo** *se e somente se*

$\forall a, b, c \in R,\ a \blacktriangledown (b \blacktriangle c) = (a \blacktriangledown b) \blacktriangle (a \blacktriangledown c)$ *e*

$a \blacktriangle (b \blacktriangledown c) = (a \blacktriangle b) \blacktriangledown (a \blacktriangle c)$

3.5.2 Reticulado de Retângulos de uma Relação Binária

Proposição 3.4

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

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

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

"$\leq\leq$" *é uma relação de ordem (parcial).*

Teorema 3.1

Seja R *uma relação binária definida sobre um conjunto* E*, e* R_R *o conjunto dos retângulos de* R *ordenados pela relação "*$\leq\leq$*". $(R_R, \leq\leq)$ é um* ***reticulado distributivo*** *com um menor elemento* $(\varnothing, E)$ *e um maior elemento* $(E, \varnothing)$ *e onde:*

$$\blacktriangledown_{j\in J}(A_j, B_j) = (\bigcup_{j\in J} A_j, \bigcap_{j\in J} B_j),$$

$$\blacktriangle_{j\in J}(A_j, B_j) = (\bigcap_{j\in J} A_j, \bigcup_{j\in J} B_j)$$

3.5.3 Reticulado de Retângulos Máximos de uma Relação Binária

Seja R uma relação binária sobre E. Para dois conjuntos A e B tal que $A \subseteq E$ e $B \subseteq E$, define-se os operandos ► e ◄ da seguinte maneira:

$$A^{\blacktriangleright} = \{d \mid \forall g, g\in A \Rightarrow (g, d) \in R\}$$

e

$$B^{\blacktriangleleft} = \{g \mid \forall d, d\in B \Rightarrow (g, d) \in R\}$$

Pode-se constatar que $A \times A^{\blacktriangleright}$ é a maior relação da forma $A \times X \subseteq R$ e que $B^{\blacktriangleright} \times B$ é a maior relação da forma $X \times B \subseteq R$. Em outras palavras, ► calcula o codomínio máximo para um domínio A e ◄ calcula o domínio máximo para um codomínio B. Os operadores ► e ◄ são antítonos na medida em que a um pequeno domínio corresponde um grande codomínio e vice-versa.

Seja a relação de ordem "$\leq\leq^{max}$", idêntica àquela que foi definida na Proposição 3.4 sobre o reticulado R_R de retângulos R (isto é, "$\leq\leq$"), mas definido sobre o subconjunto R_{Rmax} de retângulos máximos de R.

Teorema 3.2

Seja R *uma relação binária definida sobre um conjunto* E *e* R_{Rmax} *o conjunto dos retângulos máximos de* R *ordenado pela relação "*$\leq\leq^{max}$*".* (R_{Rmax}, $\leq\leq^{max}$) *é um reticulado completo onde o supremo* ($\blacktriangledown^{max}$) *e o ínfimo* ($\blacktriangle^{max}$) *de um conjunto qualquer de retângulos máximos de* R_{Rmax} *são dados respectivamente por:*

$$\blacktriangledown^{max}{}_{j\in J}(A_j, B_j) = ((\bigcup_{j\in J} A_j)^{\blacktriangleright\blacktriangleleft}, \bigcap_{j\in J} B_j),$$

$$\blacktriangle^{max}{}_{j\in J}(A_j, B_j) = (\bigcap_{j\in J} A_j, (\bigcup_{j\in J} B_j)^{\blacktriangleright\blacktriangleleft}).$$

No Anexo C é apresentado o algoritmo para construção de um grafo de retângulos ótimos a partir de uma lista de retângulos. Esta lista é constituída de vários níveis. Cada um desses níveis contém os retângulos ótimos cujo domínio ou codomínio possuem a mesma cardinalidade.

3.6 Conexão de Galois

A noção de **Conexão de Galois** será utilizada como ferramenta matemática para explorar o contexto do reticulado de retângulos máximos ou mesmo no grafo de retângulos ótimos para a formulação de uma consulta a uma base de documentos.

Notação

Se A *e* B *são dois conjuntos quaisquer, denota-se por* P(A) *e* P(B) *os conjuntos das partes de* A *e* B, *respectivamente.*

Definição 3.14

Sejam A *e* B *dois conjuntos quaisquer e as funções* σ: P(A) $\rightarrow$ P(B) *e* τ: P(B) $\rightarrow$ P(A). *O par de funções* (σ, τ) *é chamado de* **Conexão de Galois** *entre* P(A) *e* P(B) *se:*

$$\forall A' \in P(A) \text{ e } \forall B' \in P(B),\ A' \subseteq \tau(B) \Leftrightarrow B' \subseteq \sigma(A)$$

A noção de Conexão de Galois é ilustrada pela Figura 3.7.

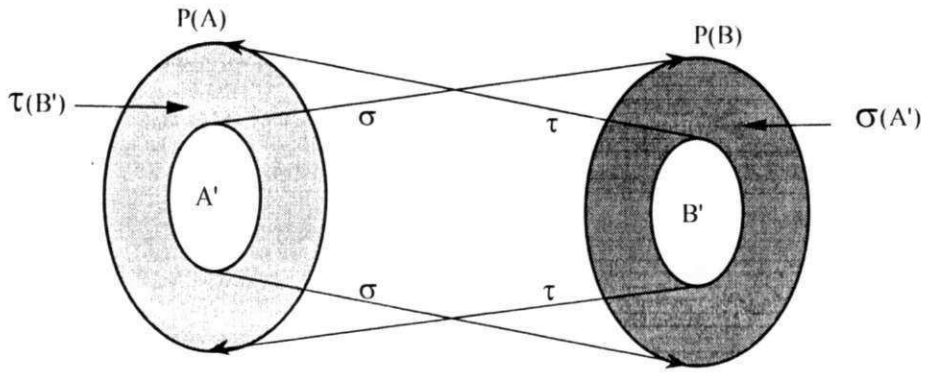


Figura 3.7: Conexão de Galois

Proposição 3.5

As funções $^{\blacktriangleright}$*: P(E)* $\rightarrow$ *P(E) e* $^{\blacktriangleleft}$*: P(E)* $\rightarrow$ *P(E) formam uma conexão de Galois.*

Observações:

1. Os operadores da Conexão de Galois ($^{\blacktriangleright}$, $^{\blacktriangleleft}$), podem igualmente ser expressos da seguinte maneira:

$$A^{\blacktriangleright} = \bigcap_{g \in A} \{g\}^{\blacktriangleright} \; e \; B^{\blacktriangleleft} = \bigcap_{d \in B} \{d\}^{\blacktriangleleft}$$

   Em particular, se (A, B) é um retângulo, então $A = B^{\blacktriangleleft}$ e $B = A^{\blacktriangleright}$.

2. Nota-se que $\{g\}^{\blacktriangleright}$ é equivalente a g.R e que $\{d\}^{\blacktriangleleft}$ é equivalente a $d.R^{-1}$.

Exemplo:

Seja RE = (A, B) *um retângulo de* R, *definido pelo conjunto* { (x, a) (x, b), (y, a), (y, b), (z, a), (z, b) }.

$$A^{\blacktriangleright} = \bigcap_{g \in A} \{g\}^{\blacktriangleright} = x.R = \{a, b\}, \; y.R = \{a, b\}, \; z.R = \{a, b\};$$

$$\bigcap (x.R, y.R, z.R) = \{a, b\} = B;$$

da mesma maneira tem-se:

$$B^{\blacktriangleleft} = \bigcap_{g \in B} \{d\}^{\blacktriangleleft} = x.R^{-1} = \{x, y, z\}, \; b.R^{-1} = \{x, y, z\} = A.$$

4. Construção Automática de um Thesaurus Retangular

A eficiência de um sistema documental está ligada à eficiência de seu mecanismo de indexação e de pesquisa. Uma etapa preliminar à geração de um *thesaurus* é a extração de palavras significativas dos documentos. No Capítulo 2 foi proposto um método automático para extração de termos.

A partir do conjunto T de termos extraídos dos documentos, é possível construir um relação binária R que associa cada elemento de T a uma lista de termos também pertencentes ao conjunto T. A partir dessa relação binária faz-se uma decomposição em um número mínimo de retângulos ótimos. Através de uma relação de ordem é construído a estrutura de reticulado do *Thesaurus* Retangular.

Este capítulo define o conceito de *Thesaurus* Retangular e os tipos de ligações semânticas existentes entre seus termos. Descreve também os passos necessários para a construção de um *Thesaurus* Retangular.

4.1 *Thesaurus* Retangular

A noção de *Thesaurus Retangular* pode ser definida como:

Definição 4.1

Um **Thesaurus Retangular** *é um grafo onde os nós são retângulos ótimos e as ligações semânticas são aquelas existentes entre os termos de um mesmo retângulo ou entre retângulos (ligações verticais e horizontais).*

As ligações verticais são definidas pela relação de ordem parcial, definida na Seção 3.5.1 (Capítulo 3).

Os nós do *Thesaurus* Retangular são os retângulos ótimos obtidos através da Decomposição Retangular de uma Relação Binária [Gammoudi 1993] expressas sob forma de uma matriz Termo-Termo.

4.1.1 Ligações Semânticas

Em um *Thesaurus* Retangular podemos distinguir quatro tipos de ligações: sinônimos, pseudo-sinônimos, ligações hierárquicas (verticais) e ligações de vizinhança (horizontais).

4.1.1.1 Sinônimos

A ligação do tipo ***sinônimo*** é uma relação de equivalência entre termos de indexação. Supõe-se que o conjunto de termos que pertencem aos domínios ou aos codomínios dos retângulos, são elementos do conjunto quociente T_p. Cada termo deste conjunto representa uma classe de equivalência. A seguir será definida a relação do tipo sinônimo entre os termos de uma mesma classe.

- Tp: conjunto dos termos que são o domínio de um retângulo ótimo qualquer
- Si: é uma relação em Tp onde $\forall\, t_i \in Tp$, $\forall\, x_i, x_j \in t_i$, x_i Si x_j

Verifica-se que **Si** é uma relação de equivalência.

- Reflexividade: $\forall\, t_i \in T_p$, $\forall\, x_i \in t_i$, x_i Si x_i
- Simetria: $\forall\, t_i \in T_p$, $\forall\, x_i, x_j \in t_i$, se x_i Si x_j então x_j Si x_i
- Transitividade: $\forall\, t_i \in T_p$, $\forall\, x_i, x_j, x_k \in t_i$, se x_i Si x_j e x_j Si x_k então x_i Si x_k

Exemplo

Os termos sinônimos de "informação" são: "fatos", "dados" (Figura 4.1).

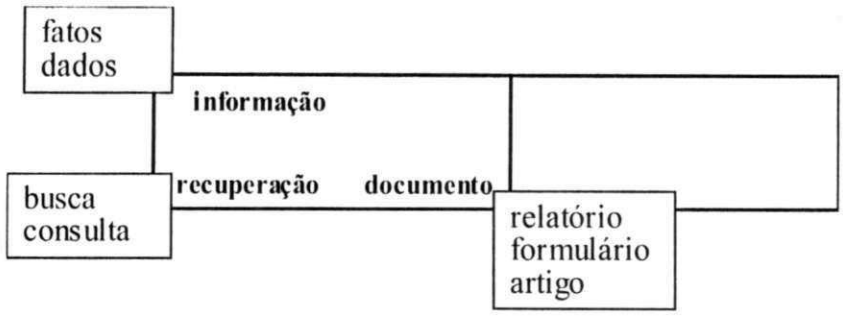


Figura 4.1: Sinônimos

Diz-se que "informação" é o representante dos seus sinônimos no retângulo onde ele aparece.

4.1.1.2 Pseudo-Sinônimos

Pseudo-sinônimos referem-se aos termos do domínio ou do codomínio de um determinado retângulo. Termos são pseudo-sinônimos se eles indexam exatamente os mesmos documentos.

Exemplo

Os pseudo-sinônimos do termo "informação" são "recuperação" e "documento" se, após a simplificação do thesaurus, *tomar-se como descritor a palavra chave "informação".*

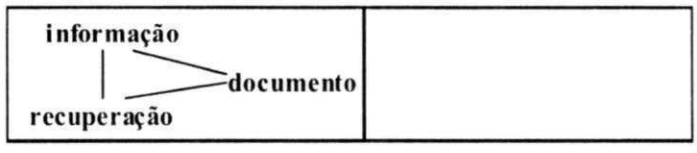


Figura 4.2: Pseudo-sinônimos

4.1.1.3 Ligações Hierárquicas

Ligações hierárquicas são definidas introduzindo-se a noção de generalização e especificação entre retângulos.

Definição: 4.2

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:*

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

Exemplo

Na Figura 4.3, o retângulo que contém os termos "informação", "recuperação" *e* "documento" *é mais específico do que os retângulos que contém respectivamente* {"informação", "documentos"} *e* {"informação", "recuperação"}.

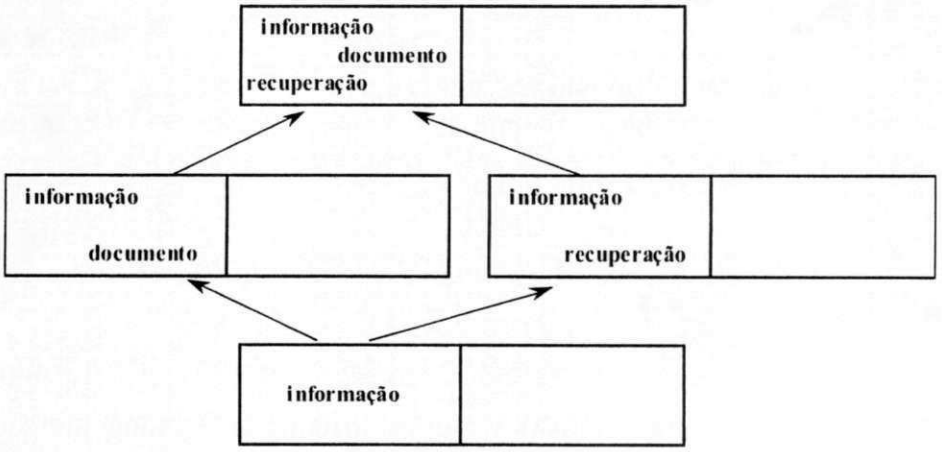


Figura 4.3: Ligações hierárquicas

É introduzida a noção de grau de generalidade/especificidade entre os retângulos para fornecer aos usuários uma ferramenta de auxílio durante a recuperação de informações.

Sejam $RE_i = (A_i, B_i)$ e $RE_j = (A_j, B_j)$ dois retângulos ótimos tal que existe uma ligação hierárquica entre RE_i e RE_j, ou seja, $card(A_i) \neq card(A_j)$, O grau de especificidade/generalidade entre retângulos ótimos de $R_{Rótimo}$ é definido como segue:

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

$$G_{g,e}(RE_i, RE_j) = \frac{1}{ABS(card(A_i) - card(A_j))}$$

Um usuário pode assim precisar o grau de proximidade entre os documentos que são recuperados.

Exemplo

Na Figura 4.3, o grau de proximidade entre o retângulo RE_1 *com domínio* A_1={"informação"} *e o retângulo* RE_2 *com domínio* A_2={"informação", "documento", "recuperação"} *é:*

$$G_{g,e}(RE_1, RE_2) = \frac{1}{ABS(1-3)} = \mathbf{0,5}$$

Exemplo

Na Figura 4.3, a proximidade entre o retângulo RE_1 *com domínio* A_1={"informação", "documento"} *e o retângulo* RE_2 *com domínio* A_2={"informação", "recuperação"} é *igual a* 1 / 3.

As ligações de vizinhança são geradas de modo dinâmico. É apresentado abaixo o algoritmo que permite recuperar a lista dos retângulos vizinhos a um retângulo dado.

4.2 Construção do *Thesaurus*

A construção de um *Thesaurus* Retangular é feita em quatro passos:

- Extração de termos e geração de uma matriz termo-termo (Capítulo 2);
- Geração dos nós do *thesaurus* (retângulos ótimos); Consiste em aplicar o método de decomposição retangular sobre a matriz binária Termo-Termo;
- Construção do grafo de retângulos ótimos;
- Simplificação do grafo de retângulos ótimos e obtenção do *thesaurus* retangular ótimo.

A denominação *Thesaurus* Retangular Ótimo é introduzida porque os nós do grafo são retângulos ótimos e porque o grafo é submetido a uma etapa de simplificação.

4.2.1 Geração da Matriz Binária Termo-Termo

A matriz binária termo-termo é gerada a partir dos cliques obtidos pelo método de extração de termos proposto em [Bruandet 1989b] e apresentado no Capítulo 2. Será utilizado o mesmo exemplo do Capítulo 2, cujos cliques são mostrados abaixo:

```
{ p y }
{ t x y z }
{ t u w }
```

A matriz binária termo-termo pode ser gerada através da combinação dois a dois dos elementos de cada clique, ou simplesmente utilizando a mesma matriz construída durante a extração dos cliques (Figura 2.3, Capítulo 2). Os pares apresentados abaixo foram gerados utilizando a primeira opção.

(p, y)	(t, x)	(t, u)
	(t, y)	(t, w)
	(t, z)	(u, w)
	(x, y)	
	(x, z)	
	(y, z)	

A matriz binária termo-termo é construída de tal forma que cada coeficiente da matriz de coordenadas (i, j), com $i \neq j$, recebe o valor 1 se existe uma ligação entre os termos t_i e t_j, e recebe o valor 0, caso contrário.

Termo	p	t	u	w	x	y	z
p	-	0	0	0	0	1	0
t	-	-	1	1	1	1	1
u	-	-	-	1	0	0	0
w	-	-	-	-	0	0	0
x	-	-	-	-	-	1	1
y	-	-	-	-	-	-	1
z	-	-	-	-	-	-	-

Figura 4.5: Matriz binária termo-termo

A utilização dos cliques como fonte de informação para a construção da matriz termo-termo é opcional, podendo-se utilizar a mesma matriz construída na fase de extração de termos (cliques).

4.2.2 Geração dos Nós

Utilizando o algoritmo da Decomposição Retangular sobre a matriz binaria termo-termo [Gammoudi 1993] (Anexo C), obtém-se os seguintes retângulos ótimos:

$$\{ p, t, x \} \times \{ y \}$$

$$(t, x, y \} \times \{ z \}$$

$$\{ t, x \} \times \{ y, z \}$$

$$\{ t, u \} \times \{ w \}$$

$$\{ t \} \times \{ x, y, z, u, w \}$$

4.2.3 Geração e Simplificação do Grafo de Retângulos

Utilizando-se o algoritmo para a geração do grafo de retângulos [Gammoudi 1993] (Anexo D) sobre os retângulos ótimos obtidos na etapa anterior, obtém-se o *thesaurus* (grafo de retângulos) "em estado bruto" (Figura 4.6).

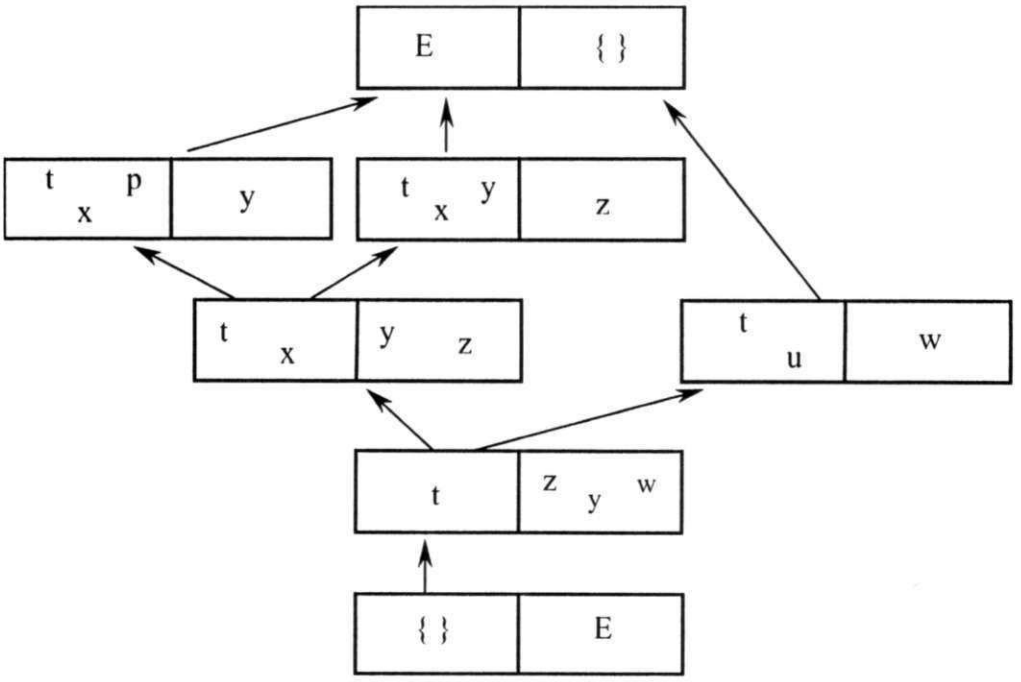


Figura 4.6: *Thesaurus* Retangular antes da simplificação

A última etapa é a simplificação do *thesaurus*. O sentido do percurso do *Thesaurus* durante o processo de simplificação não é importante (de baixo para cima ou de cima para baixo). Para cada retângulo encontrado, o sistema armazena somente o domínio (ou o codomínio). Deve ser considerado aquele de menor cardinalidade. O conjunto retido é chamado **representante** do retângulo.

Definição 4.4

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

A_i, *se* $card(A_i) \leq card(B_i)$

B_i, *se* $card(B_i) < card(A_i)$

O representante de um retângulo contém os termos com 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, mais ele veicula semântica e permite assim um maior número de termos. Verificamos que o representante de um retângulo é um subconjunto do domínio ou do codomínio de seus retângulos específicos.

O princípio da simplificação é que o representante de um retângulo é suprimido do domínio (ou do codomínio) de cada um de seus retângulos específicos.

```
Simplificacao( Re, G)
inicio
     para x ∈ pred( Re, G ) faça
     inicio
        s := representante(X)
        para y ∈ pred( X, G ) faça
        inicio
           Suprimir( s, y )
           Profundidade( s, y, G )
           Simplificacao( X, G )
        fim
     fim
fim

Profundidade(s, y, G)
inicio
     para Z ∈ pred(y, G) faça
     inicio
        Suprimir(S, Z)
        Profundidade(S, Z, G)
     fim
fim
```

Após a aplicação do algoritmo sobre o grafo, obtemos o seguinte grafo simplificado.

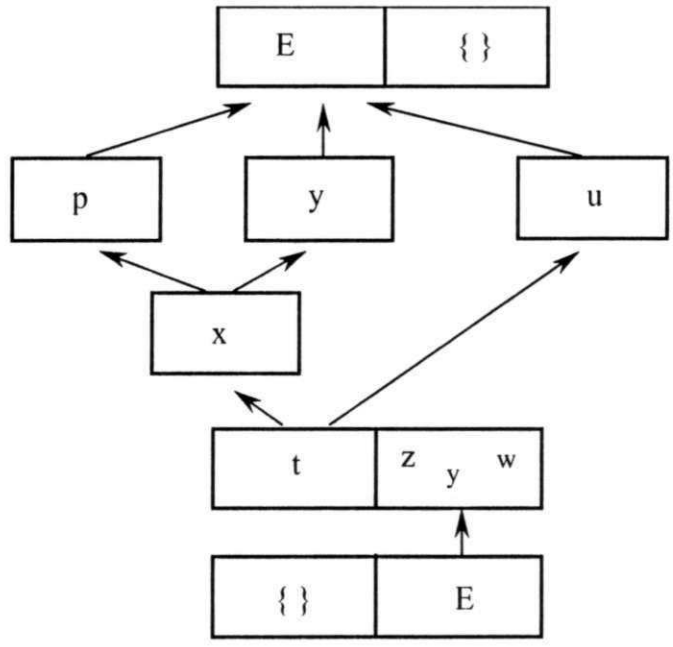


Figura 4.7: *Thesaurus* Retangular "ótimo"

Após a simplificação, é obtido um grafo cujos nós são termos representantes dos retângulos. Esse novo grafo é chamado ***Thesaurus* Retangular Ótimo** (Figura 4.7).

5. Organização Hierárquica dos Documentos

O fato da definição dos diferentes componentes de um sistema documental estar baseado em um mesmo método contribui para a melhoria de seu desempenho, ou pelo menos determina uma uniformidade conceitual no sistema. Partindo-se desse princípio, a representação da semântica de uma base documental é feita utilizando o mesmo método utilizado para a construção do *Thesaurus* Retangular vista no capítulo anterior.

Seja D o conjunto de documentos do *thesaurus* e T o conjunto de termos extraídos de D. Representamos a semântica de uma base documental como uma relação binária R cujo domínio é o conjunto T e o codomínio é o conjunto D. Utiliza-se o método de *Decomposição Retangular de uma Relação Binária* [Gammoudi 1993] para classificar automaticamente os documentos ao mesmo tempo que seus descritores. Aplicando-se a mesma relação de ordem utilizada durante a construção do *Thesaurus* Retangular, as classes (retângulos ótimos) obtidas vão ser organizadas na forma de um grafo de retângulos ótimos.

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

- Construção da matriz binária Termo-Documento representando a semântica da base de documentos;

- Agrupamento dos documentos e seus descritores na forma de retângulos ótimos (classes) através da decomposição da matriz Termo-Documento;
- Organização dos diferentes retângulos, obtidos com a etapa anterior, na forma de um grafo de retângulos.

5.1 Geração da matriz Termo-Documento

Durante a extração do conjunto dos termos pelo método de Bruandet [Bruandet 1989b] supõe-se conhecer o conjunto dos documentos indexados por esses termos. A relação binária expressa pela matriz Termo-Documento é portanto deduzida pela correspondência entre o conjunto dos termos e o conjunto dos documentos.

5.2 Classificação Automática dos Documentos

Organiza-se as diferentes classes (retângulos ótimos) de documentos na forma de um grafo de retângulos.

Seja R uma relação binária que representa a semântica de uma Base de Documentos definida de T em D, onde T é o conjunto dos termos que indexam os documentos D (Figura 5.1).

T \ D	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11
t1		1	1		1						
t2		1	1								
t3		1	1								
t4		1	1			1					
t5		1	1		1	1	1	1			
t6							1	1			
t7				1							
t8				1					1		
t9				1					1	1	
t10				1							
t11											1
t12			1	1							1

Figura 5.1: Relação inicial R

A decomposição de R nos dá o conjunto dos retângulos ótimos da Figura 5.2.

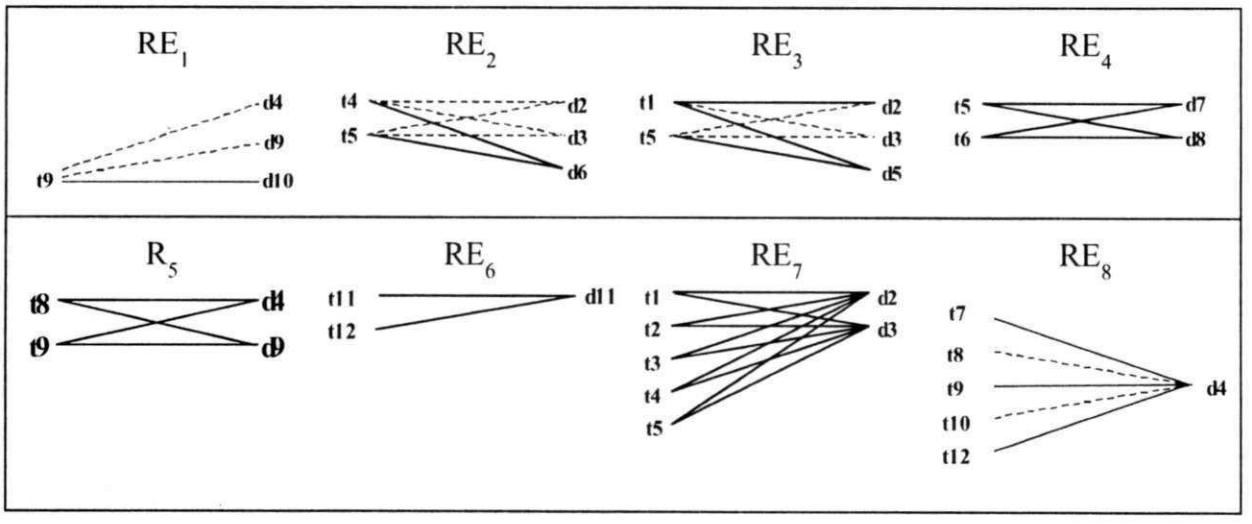


Figura 5.2: Cobertura de R por 8 retângulos

Após eliminação dos elementos redundantes, obtém-se a cobertura de R. A Figura 5.3 mostra uma representação da base de documentos na forma de um grafo de retângulos ótimos.

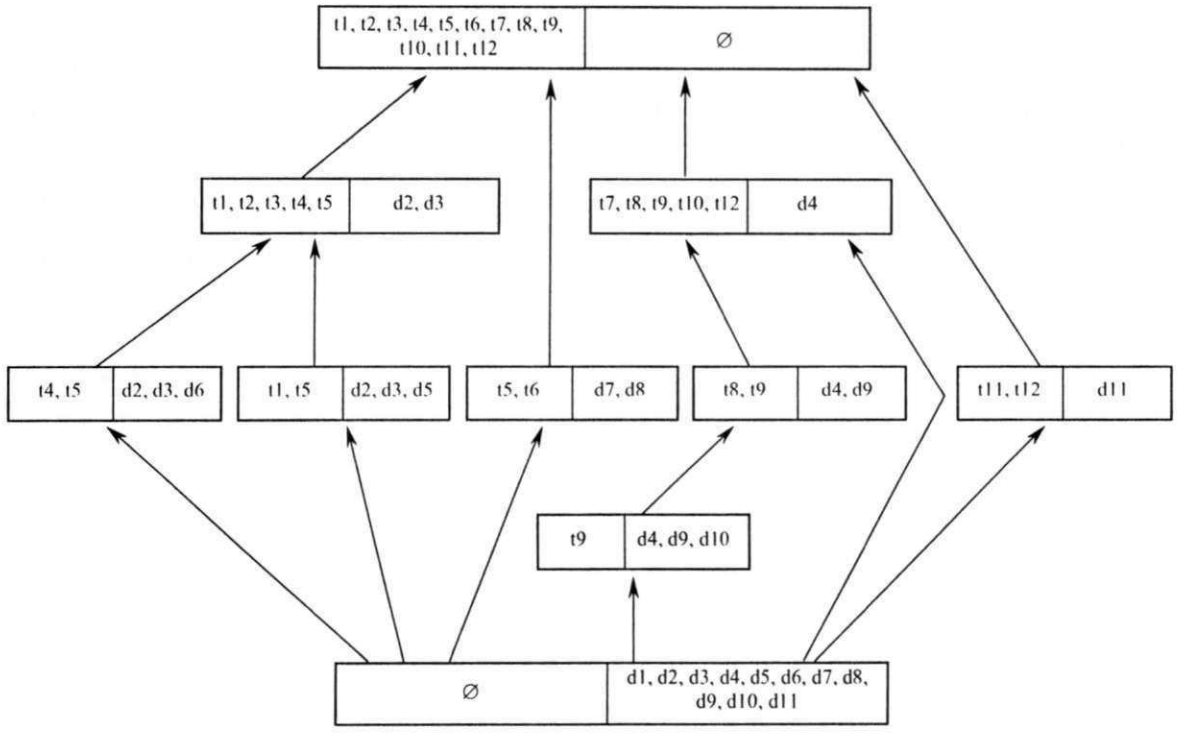


Figura 5.3: Organização hierárquica da base de documentos

Observação: $(A_1, B_1) \rightarrow (A_2, B_2) \Rightarrow (A_1, B_1) \leq\leq^{\text{ótimo}} (A_2, B_2)$

6. Métodos de Pesquisa em um Sistema Documental Retangular

Neste capítulo serão detalhadas as diferentes operações de pesquisa de acordo com a maneira pela qual uma consulta é expressa. A noção de Conexão de Galois será utilizada como ferramenta matemática para explorar o contexto do grafo de retângulos ótimos para a formulação de uma consulta a uma base de documentos. Será mostrado como, graças às conexões de Galois, é possível reformular automaticamente as consultas realizadas pelos usuários, a fim de auxiliar os usuários iniciantes.

O processo de pesquisa em um sistema documental retangular consiste de 5 operações: *Formulação da consulta*, *Tradução da consulta em um sistema de inequações*, *Execução da consulta*, *Visualização da resposta* e *Reformulação da consulta*.

6.1 Formulação de uma Consulta

Uma consulta pode ser uma expressão booleana, uma expressão em linguagem natural ou pode ainda ser construída a partir de uma seqüência de seleções de menus ou nós em um grafo.

Uma consulta documental poderá ser expressa de duas maneiras. Uma consulta poderá ser feita através de uma expressão em linguagem natural ou a partir da navegação no

thesaurus retangular (seleção de nós). Uma vez definida a consulta, a etapa seguinte consiste em traduzi-la em uma expressão intermediária a fim de que ela possa ser executada.

6.2 Tradução da Consulta em um Sistema de Inequações

Essa operação consiste em traduzir a consulta em uma ou mais inequações do tipo $T \times D \subseteq R(R)$, onde:

- T é conjunto de termos extraídos da consulta;
- D é o conjunto de documentos indexados pelos termos de T;
- $T \times D$ é o retângulo a ser procurado;
- R(R) é o grafo de retângulos que representa a base de documentos.

6.3 Execução e Visualização de uma Consulta

A execução de uma consulta é obtida pela resolução de uma ou várias inequações. Se a consulta é traduzida na forma de várias inequações, essas podem ser executadas em paralelo, o que permite um ganho no tempo de execução. A solução dessas inequações constitui o conjunto dos documentos procurados.

A visualização da resposta a uma consulta consiste em visualizar os documentos encontrados utilizando-se uma ferramenta de apresentação.

6.4 Reformulação de uma consulta

Quando o usuário não está satisfeito com a resposta apresentada, é desejável a possibilidade da reformulação da consulta. Serão utilizadas as conexões de Galois para auxiliar essa reformulação. Na Seção 6.5.5 deste capítulo, mostraremos como esta reformulação pode ser feita utilizando a conexão de Galois.

6.5 Ilustração de uma Pesquisa

A seguir serão apresentadas duas sessões de pesquisa em um sistema documental retangular. Na primeira sessão tratamos o caso onde a consulta é expressa a partir do *thesaurus* retangular e na segunda sessão será apresentado um caso onde a consulta é expressa em linguagem natural.

6.5.1 Formulação da consulta a partir do *Thesaurus*

Pouco importa o tipo de usuário (leigo ou experiente), a expressão da consulta a partir do *thesaurus* permite fornecer-lhe os diferentes tipos de relações semânticas entre os termos. Assim, graças às relações hierárquicas, o usuário pode selecionar a partir do *thesaurus*, termos que podem ser mais genéricos ou mais específicos. Ele pode também utilizar termos sinônimos ou ainda expressar sua consulta por termos que abrangem domínios eventualmente diferentes por meio das relações de vizinhança.

6.5.2 Formulação da consulta em linguagem natural

Essa forma de expressar uma consulta é mais "confortável" para o usuário, mas também mais complexa de ser tratada. É proposto um método idêntico ao que foi utilizado para a indexação dos documentos. A consulta sofre as duas operações seguintes:

- Extração dos conceitos da consulta de forma análoga à extração de termos apresentado no Capítulo 2 e obtenção do grafo G_c que a representa (Figura 6.1).
- Execução da consulta resolvendo a inequação $A \times X \subseteq R(R)$, onde A representa o conjunto de termos (Nós) de G_c.

Observação

Os conjuntos **X**, **S** e **A**, que serão utilizados na seqüência deste capítulo, estão representados graficamente na Figura 6.2.

6.5.3 Tradução da consulta

Seja $(R_{ótimo}, \leq\leq^{ótimo})$ o grafo dos retângulos ótimos de uma relação binária **R** definida sobre um conjunto **E**. A resolução da inequação

$$\exists ?\, X \subseteq E \mid S \times X \subseteq R \; e \; A \subseteq S$$

que chamamos "inequação retangular", se reduz a busca do retângulo

$$\mathbf{e} \in (R_{ótimo}, \leq\leq^{ótimo}) \mid \mathbf{e} = S \times X \; e \; A \subseteq S$$

Toda consulta deve ser traduzida para essa forma antes de ser executada.

Exemplo

Seja C *uma consulta expressa em linguagem natural e* Gc *o grafo associado obtido pelo método de* Bruandet . *O* clique *correspondente a* Gc *é* (t_1, t_2, t_3, t_4) *(Figura 6.1)*

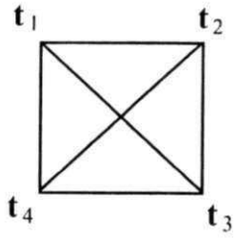


Figura 6.1: Grafo G_c representando uma consulta C

O conjunto { t_1, t_2, t_3, t_4 } constitui portanto o domínio do retângulo procurado no grafo que representa a base de documentos. A consulta pode portanto ser expressa como:

$$\exists ?\ \{ t_1, t_2, t_3, t_4 \} \subseteq E \mid \{t_1, t_2, t_3, t_4 \} \times X \subseteq R(R)$$

6.5.4 Execução da consulta

Nessa seção baseamo-nos em exemplos para explicar os passos de execução de uma consulta. Os exemplos aqui apresentados baseiam-se na Figura 5.3, Capítulo 5.

Exemplo: Consulta conjuntiva

"Quais são todos os documentos indexados pelos termos t_4 e t_5 (isto é, que tratam ao mesmo tempo de t_4 e t_5) ?"

Para isso é preciso encontrar o retângulo RE *no grafo* R(R) *tal que* RE = A×X, *onde o desconhecido é* X *(documentos procurados) e onde* A = $\{t_4, t_5\}$. *Em outras palavras, trata-se de encontrar o retângulo* RE *cujo domínio é* A *(ou contém* A*) e cujo codomínio é o maior possível. Para isso utilizamos a função* ► *para determinar* $\{t_4, t_5\}^{\blacktriangleright}$. $\{t_4, t_5\}^{\blacktriangleright}$ = $\{d_2, d_3, d_6\}$. *Esse resultado corresponde ao retângulo ótimo* RE = (A, B) *com* A = $\{t_4, t_5\}$ e B = $\{d_2, d_3, d_6\}$. *Podemos então deduzir que o conjunto procurado dos documentos que satisfazem a consulta é* X=B=$\{d_2, d_3, d_6\}$.

Exemplo: Consulta disjuntiva

Quais são todos os documentos indexados pelos termos (t_1 e t_5) ou pelos termos (t_8 e t_9) ?

O tratamento dessa consulta se reduz a resolução de duas inequações retangulares

$$\exists? X_1 \subseteq D \mid A_1 \times X_1 \subseteq R \text{ com } A_1 = \{t_1, t_5\} \qquad \text{(i)}$$

$$\exists? X_2 \subseteq D \mid A_2 \times X_2 \subseteq R \text{ com } A_2 = \{t_8, t_9\} \qquad \text{(ii)}$$

Aplicando-se sucessivamente sobre cada uma dessas duas inequações retangulares o mesmo raciocínio do exemplo anterior, encontramos a solução à inequação (**i**)*:* $X_1 = \{d_2, d_3, d_5\}$*, e a solução da inequação* (**ii**)*:* $X_2 = \{d_4, d_9\}$*. Consequentemente, a solução global do conjunto de inequações acima é* $X = X_1 \cup X_2$*. Assim a resposta à consulta é o conjunto de documentos* $X = \{d_2, d_3, d_4, d_5, d_9\}$.

6.5.5 Reformulação da consulta

Através de dois exemplos será explicado como é realizada a reformulação automática de uma consulta, através das conexões de Galois. ($^{\blacktriangleright}$ $^{\blacktriangleleft}$).

Exemplo

Quais são os documentos indexados pelos termos t_2 e t_3 e t_4 ?

O resultado a essa consulta é $\{t_2, t_3, t_4\}^{\blacktriangleright} = \{d_2, d_3\}$*. Aplicamos a esse resultado a função* $^{\blacktriangleleft}$ *para fornecer, caso existam, os termos de indexação suplementares que permitem ter a mesma solução.*

Assim o usuário pode enriquecer sua consulta e ter uma idéia mais clara sobre sua reformulação. $\{t_2, t_3, t_4\}^{\blacktriangleleft\blacktriangleright} = \{d_2, d_3\}^{\blacktriangleleft} = \{t_1, t_2, t_3, t_4, t_5\}$*, que corresponde ao retângulo* RE=(A, B)*, com* $A=\{t_1, t_2, t_3, t_4, t_5\}$ *e* $B=\{d_2, d_3\}$*. Podemos portanto deduzir que o conjunto dos documentos que respondem à consulta é* X=B={d2, d3}*. Entretanto constatamos nesse exemplo que* $A \subseteq \{d_2, d_3\}^{\blacktriangleleft}$ *Isso traduz o fato que os documentos que respondem à consulta são todos indexados por outros termos (aqui por* t_1 *e* t_5*) além de serem indexados pelos termos* t_2, t_3 *e* t_4.

Ou seja, a resposta à consulta

"Quais são todos os termos indexados por t_1, t_2, t_3, t_4 ***e*** t_5 ***?"***

é a mesma que a consulta inicial de forma ainda mais precisa. É portanto possível fornecer automaticamente ao usuário uma resposta "enriquecida" e que informa que os documentos que ele procura são também indexados pelos termos t_1 *e* t_5 *e que a consulta pode ser reformulada considerando essa informação.*

Seja $(G_{Rótimo}, \leq\leq^{ótimo})$ o grafo dos retângulos ótimos de uma relação binária R definida sobre um conjunto E. A inequação

$$\exists? X \subseteq E \mid S \times R \subseteq R \text{ e } A \subseteq S$$

que chamamos de "inequação retangular" se reduz à busca de um retângulo

$$\mathbf{e} \in (R_{ótimo}, \leq\leq^{ótimo}) \mid \mathbf{e} = S \times X \text{ e } A \subseteq S$$

Na Figura 6.2, temos $A^{\blacktriangleright} = X$ e $X^{\blacktriangleleft} = S$, onde A poderia ser, por exemplo, um conjunto de termos de indexação e X o conjunto dos documentos procurados e indexados pelos termos de A. Essas duas operações nos permitem:

- Obter a solução da inequação $\exists? X \subseteq E \mid S \times R \subseteq R \text{ e } A \subseteq S$
- "enriquecer" a resposta à consulta do usuário quando o retângulo encontrado, admite X como codomínio, possui um domínio S que é um sobreconjunto de A (Figura 6.2)

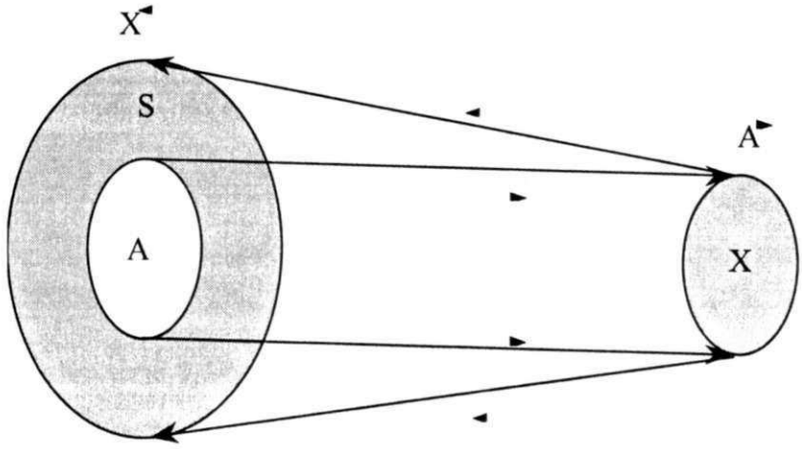


Figura 6.2: Reformulação de uma consulta através das Conexões de Galois

No caso onde o usuário não estiver satisfeito com a resposta obtida a partir de sua consulta é possível, graças às Conexões de Galois, fornecer-lhe o conjunto superior dos termos que lhe permite obter os mesmos resultados. Assim o usuário pode ter uma idéia precisa sobre o conjunto de termos que ele pode utilizar, evitando assim a redundância em sua consulta. Graças à reformulação automática, o usuário pode refinar sua consulta explorando a riqueza semântica oferecida pelo *thesaurus* retangular.

7. Implementação e Resultados experimentais

O trabalho de Marie-Françoise Bruandet [Bruandet 1989] (Capítulo 2), e o trabalho de Mohamed Gammoudi [Gammoudi 1993] (Capítulos 3 a 6), fornecem uma sólida base teórica para a implementação da aplicação que será descrita neste capítulo. O sistema foi desenvolvido utilizando a linguagem Delphi 3.0 [Henderson 1996] em ambiente Windows 95. Na Figura 7.1 é apresentada a janela de apresentação do sistema.

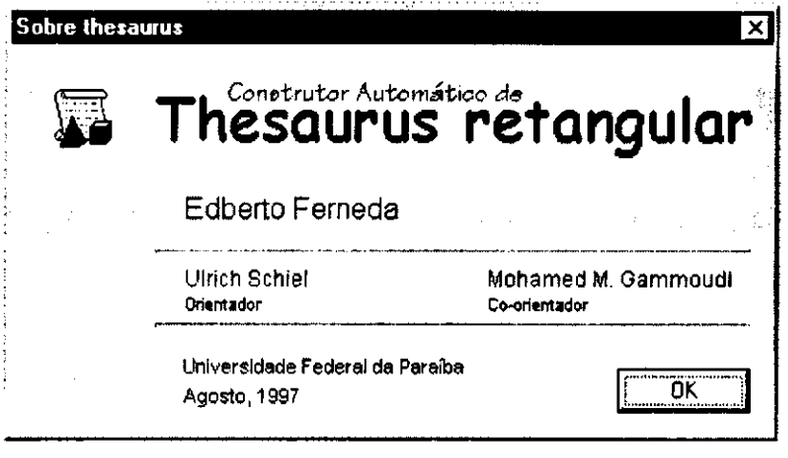


Figura 7.1: Janela de apresentação do Sistema

O sistema desenvolvido está longe de ser um produto comercialmente aceitável. Ele apenas pretende validar a maioria das idéias deste trabalho de forma concreta, utilizando uma

interface moderna e amigável, com janelas, menus, botões, e diversas características presentes nos *softwares* atuais. A Figura 7.2 apresenta a janela principal do sistema.

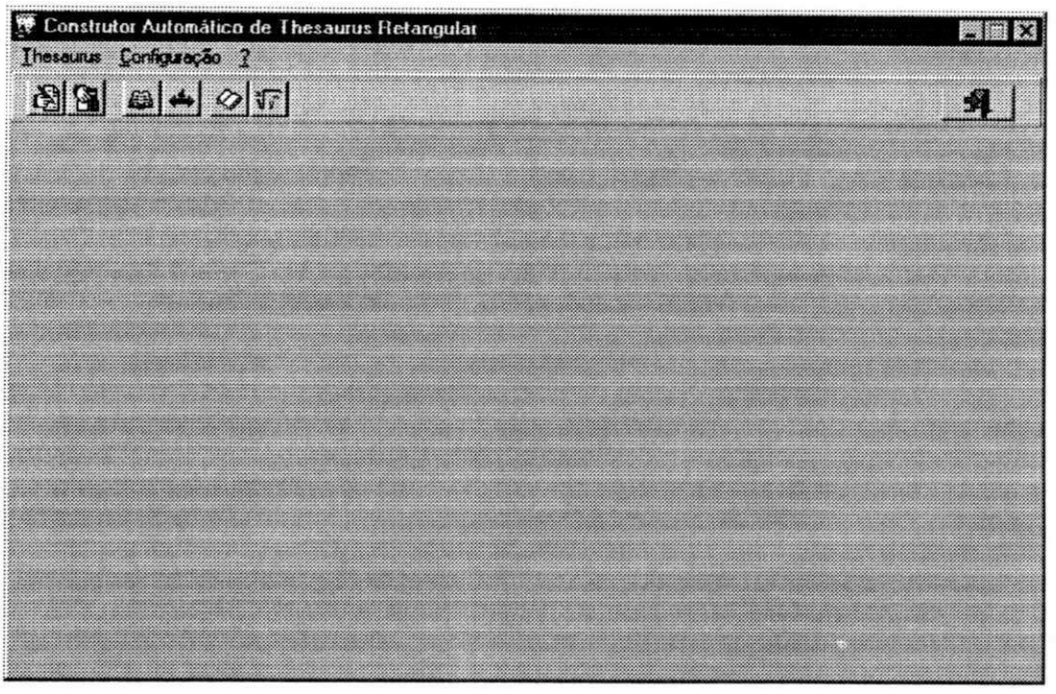


Figura 7.2: Janela principal

A seguir serão descritas as diversas funções do sistema, fazendo, sempre que possível, referência à fundamentação teórica apresentada nos capítulos anteriores.

7.1 Categorias Gramaticais

No processo de extração de termos, apresentado no Capítulo 2, é calculado, para cada par de termos, um valor que expressa a força de ligação entre eles. Para o cálculo desse valor leva-se em consideração a categoria gramatical dos termos. É necessário portanto especificar quais as categorias gramaticais que serão consideradas durante o processo de extração de termos. A Figura 7.3 apresenta a janela do sistema onde o usuário define as categorias gramaticais que julgar serem necessárias.

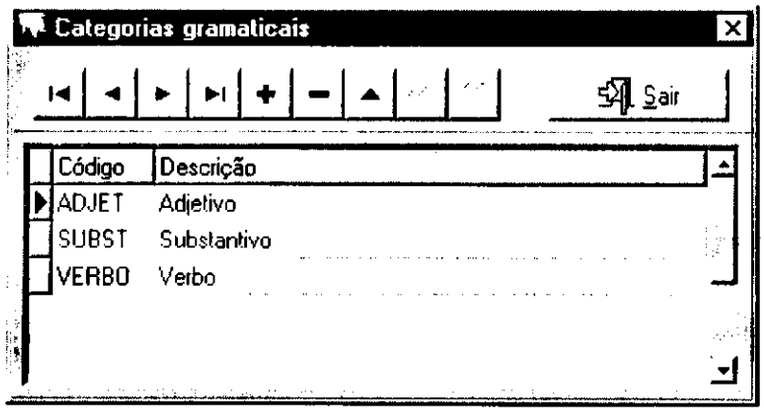


Figura 7.3: Janela para a definição das categorias gramaticais

7.2 Distância entre Categorias Gramaticais

Na equação 3 do Capítulo 2 foi introduzido um limite *t* para a distância máxima entre dois termos. Na equação 4 do mesmo capítulo, esse limite foi formalizado da seguinte maneira:

$$t(x, y) = \text{LIMITE}[\ \text{CAT}(x), \text{CAT}(y)\ ]$$

onde LIMITE *é a distância máxima entre a categoria gramatical de x,* CAT(x), *e a categoria gramatical de y,* CAT(y).

Os valores de *t* são definidos entre duas categorias gramaticais definidas pelo usuário na seção anterior (seção 7.1).

Na Figura 7.4 é apresentada janela do sistema onde o usuário define o limite *t* (**Distância**) entre duas categorias gramaticais (**Categoria 1** e **Categoria 2**).

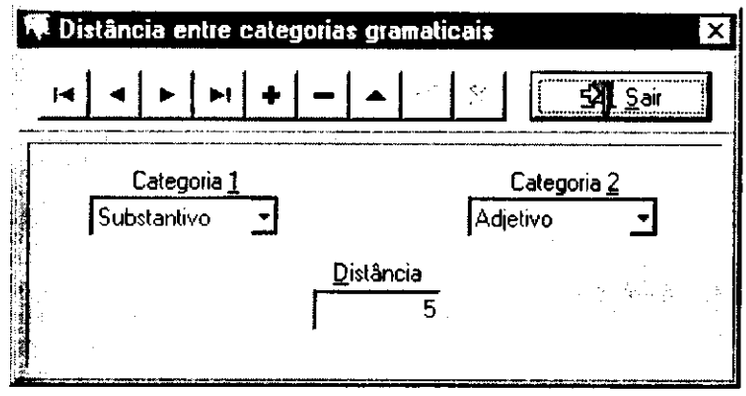


Figura 7.4: Janela para especificação da distância entre categorias gramaticais

As categorias gramaticais e as distâncias máximas entre elas podem ser representadas através de um grafo, como mostrado na Figura 2.1 do Capítulo 2.

7.3 Dicionário

Durante o processo de extração de termos, as palavras extraídas dos textos devem ser reduzidas à uma forma normalizada (termo). Além disso, é necessário obter a categoria gramatical de um termo para se obter o limite ***t***, utilizado no cálculo da força de ligação M_2 (equação 4, Capítulo 2).

O sistema se utiliza de um dicionário para obter não só a forma normalizada de uma palavra mas também a categoria gramatical a qual ela pertence. A Figura 7.5 mostra a janela do sistema onde o usuário define o dicionário, informando a palavra, sua forma normalizada (termo) e a categoria gramatical a qual pertence.

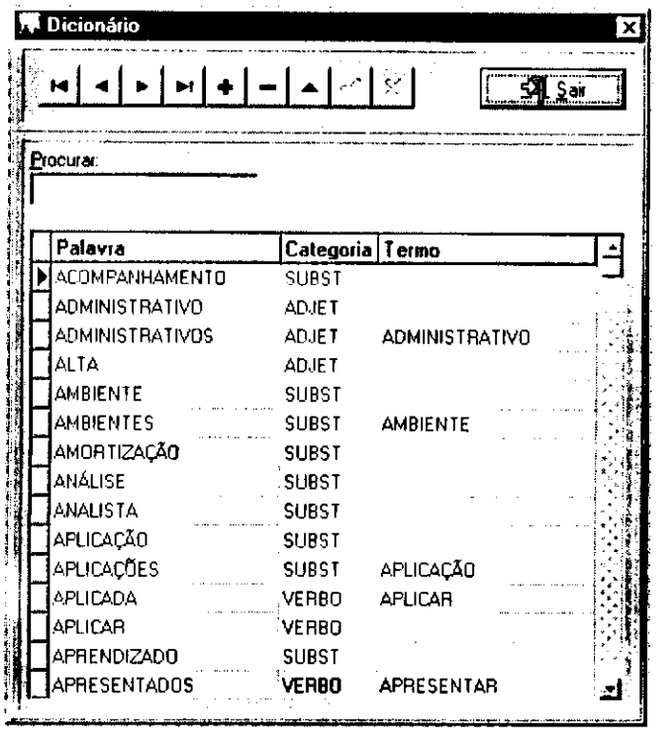


Figura 7.5: Janela para definição do Dicionário

7.4 Parâmetros

Nas equações 8 e 9 do Capítulo 2 é introduzido um fator de correção ***k*** para a força de ligação M_1 (equação 7), resultando em uma nova medida M_2. Na definição desse fator, é utilizado um parâmetro ***n*** (Força de ligação)

Um outro parâmetro é utilizado na transformação da matriz termo-termo contendo os valores da medida M_2 para a matriz binária termo-termo. Esse parâmetro, chamado **Força**

mínima de ligação é utilizado para eliminar ligações consideradas fracas (Seção 2.2 do Capítulo 2).

Através da janela apresentada na Figura 7.6 o usuário especifica o valor de ***n*** e também o valor da **Força mínima de ligação** a ser considerada na construção da matriz binária.

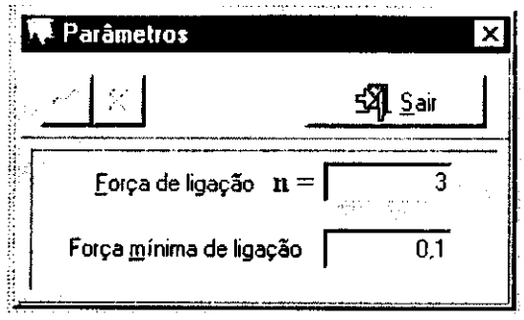


Figura 7.6: Janela para a definição de parâmetros

7.5 Definição do *Thesaurus*

Um *thesaurus* é construído a partir de um conjunto de documentos. O Sistema permite a definição de um ou mais *thesaurus* através da definição dos documentos que serão utilizados na sua construção. A Figura 7.7 mostra a janela onde o usuário define o(s) *thesaurus* e o conjunto de documentos correspondentes.

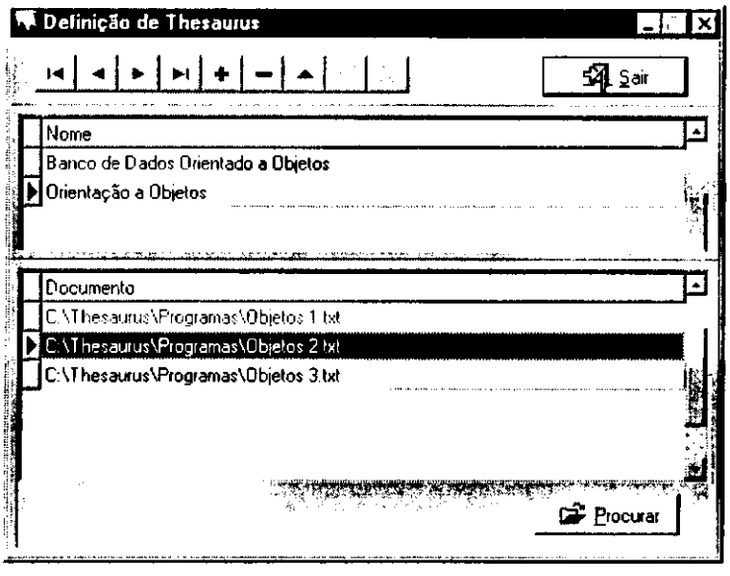


Figura 7.7: Janela para a definição de thesaurus e seus documentos

7.6 Construção do *Thesaurus*

Por se tratar de um Sistema experimental, ele foi implementado com a finalidade de apresentar os resultados das principais fases de construção do *thesaurus*. Para isso, foi

necessária a utilização de alguns componentes de interface que futuramente poderiam ser eliminados para tornar o Sistema mais rápido, robusto e comercialmente aceitável.

A janela de construção de *thesaurus* (Figura 7.8) possui 7 divisões onde são apresentadas sob forma de texto os resultados de alguns dos passos necessários para a construção do *Thesaurus Retangular*, a organização hierárquica dos documento e, por fim, é apresentada uma divisão que permite ao usuário fazer consultas ao Sistema Documental.

As divisões possuem as seguintes funcionalidades:

- Extração de termos (Capítulo 2);
- Construção da matriz binária termo-termo (Capítulo 2);
- Extração dos cliques (Capítulo 2);
- Geração do grafo de retângulos ótimos (termo-termo) (Capítulo 4);
- Construção da matriz termo-documento (Capítulo 5);
- Geração do grafo termo-documento (Capítulo 5);
- Execução de consultas (Capítulo 6);

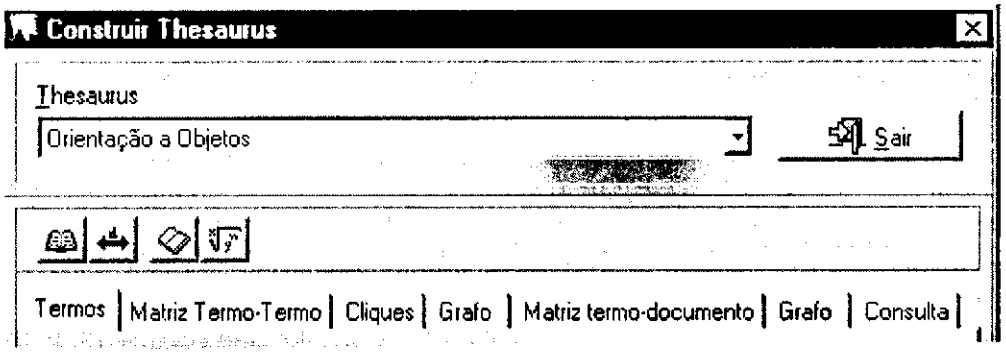


Figura 7.8: Janela de construção do *thesaurus*

Inicialmente o usuário deve especificar o *thesaurus* que se pretende construir, escolhendo um dos *thesaurus* previamente definidos (Seção 7.5). Dependendo dos resultados obtidos, o usuário poderá mudar os parâmetros utilizando um conjunto de botões que aparecem na janela.

7.6.1 Extração de Termos

A primeira fase para a construção do *thesaurus* é a extração de termos do conjunto de textos. Durante o processo de extração de termos o sistema apresenta, além do termo

propriamente dito, o seu endereço no formato <ND, NF, NT>, como especificado na equação 1 do Capítulo 2, e a sua categoria gramatical. A Figura 7.9 apresenta os termos extraídos do conjunto de textos do *thesaurus* denominado "Orientação a Objetos".

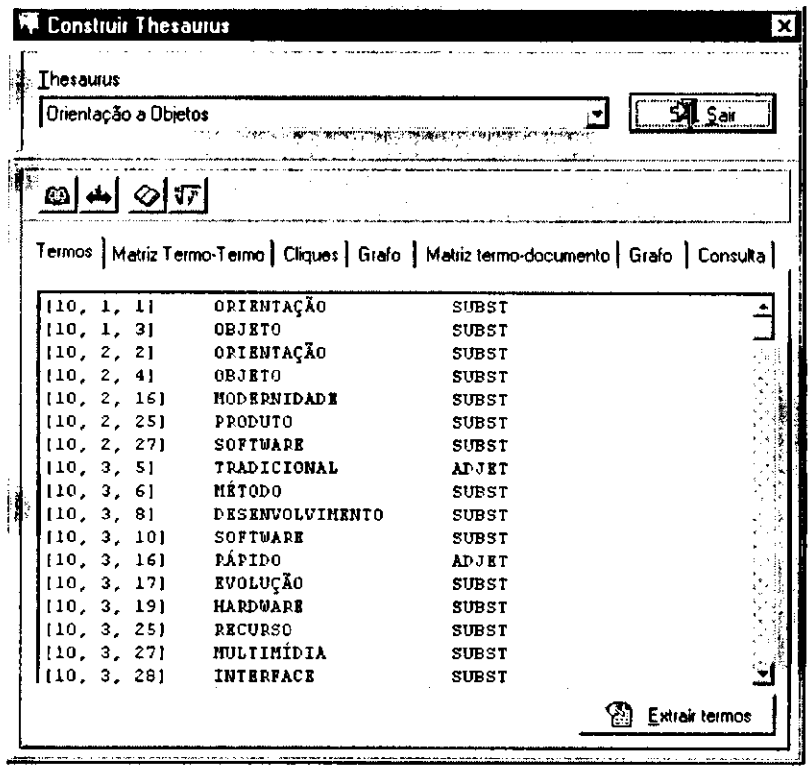


Figura 7.9: Extração de termos

7.6.2 Construção da Matriz Binária Termo-Termo

Os termos extraídos dos textos são utilizados para compor as linhas e as colunas de uma matriz (Figura 7.10). Para cada par de termos diferentes é calculado a força de ligação M_2, descritas nas equações 8 e 9 do Capítulo 2. Caso a medida M_2 seja maior ou igual à força mínima de ligação, especificada como parâmetro (seção 7.4, Figura 7.6), a célula correspondente é assinalada especificando a ligação entre os dois termos (ver seção 2.2 do Capítulo 2).

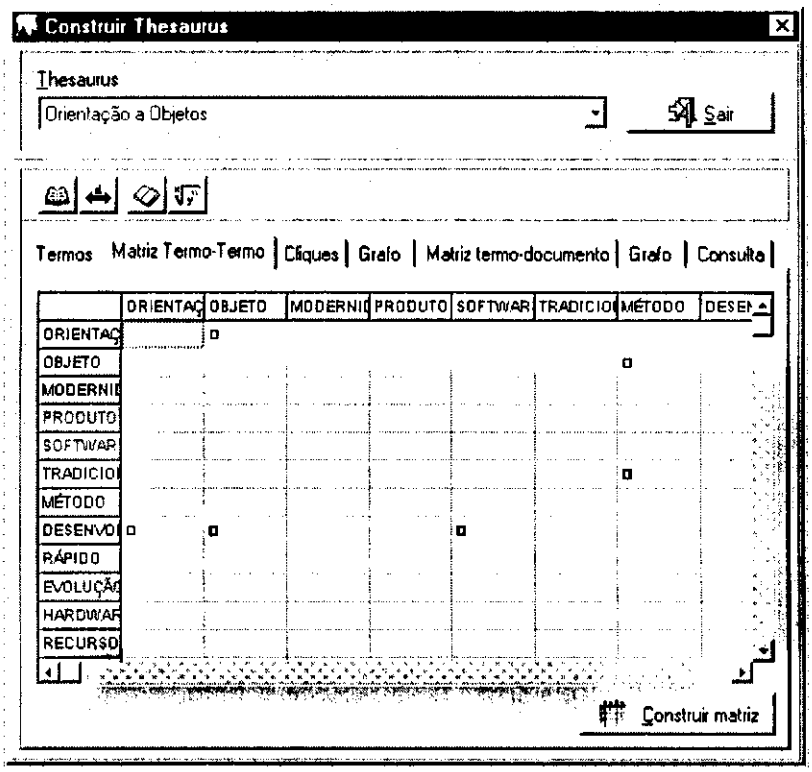


Figura 7.10: Representação da Matriz Binária Termo-Termo

7.6.3 Extração de Cliques

A partir da matriz binária termo-termo, calculada na seção anterior, extrai-se os *cliques* do grafo definido pela matriz, utilizando-se o algoritmo descrito no Anexo A.

A Figura 7.11 apresenta os cliques extraídos da matriz binária termo-termo mostrada na Figura 7.10 da seção anterior.

Os cliques são importantes para a avaliação do processo de extração de termos. Caso os resultados se mostrem insatisfatórios, é possível ajustar a aplicação através de ajustes nos parâmetros (seção 7.4) ou alterando a distância entre as categorias gramaticais (seção 7.2).

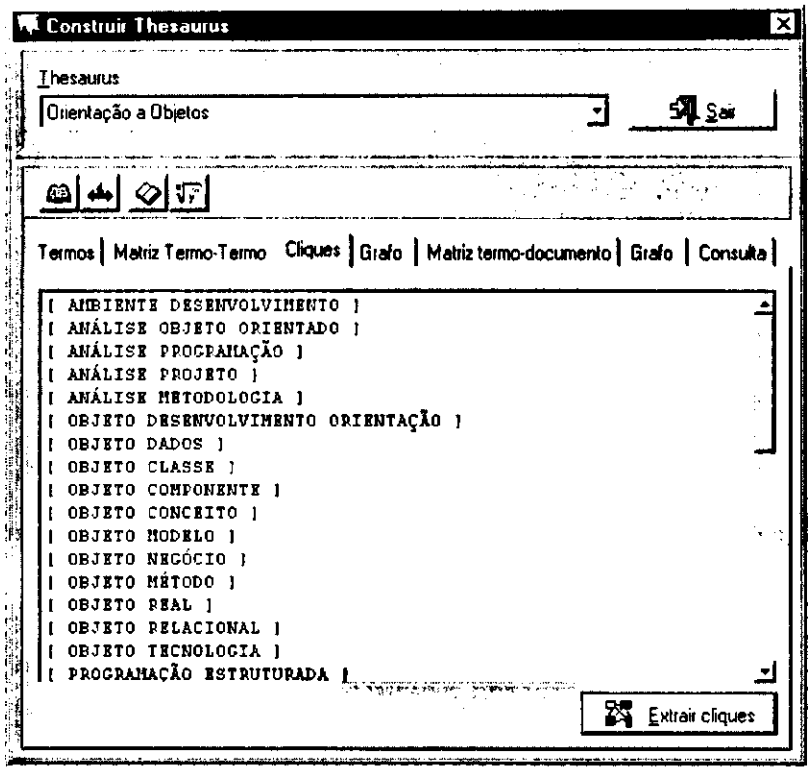


Figura 7.11: Cliques extraídos da Matriz Binária Termo-Termo

7.6.4 Construção do Grafo de Retângulos

A construção do grafo de retângulos é feita utilizando os conceitos apresentados na seção 3.5.3 (Capítulo 3) e o algoritmo do Anexo D.

A Figura 7.12 mostra a janela onde o grafo de retângulos é apresentado. A representação do grafo é feita de forma textual, sendo de difícil visualização e análise. Porém, como se trata de uma implementação experimental, optou-se por esse tipo de representação, tendo em vista a dificuldade de implementação de uma representação gráfica.

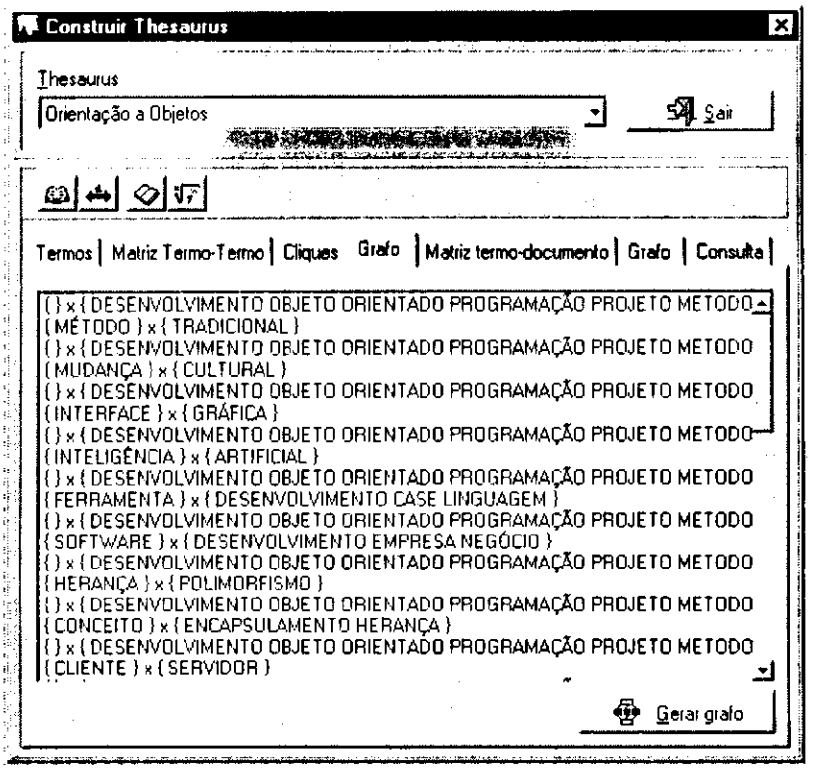


Figura 7.12: Representação textual do grafo de retângulos ótimos

7.7 Organização Hierárquica dos Documentos

A organização da base de documentos é feita utilizando a mesma fundamentação matemática utilizada para a construção do gafo de retângulos ótimos, o que garante uma uniformidade conceitual à aplicação. A diferença principal é que a relação binária utilizada para a organização hierárquica dos documentos é representada por uma matriz termo-documento (Capítulo 5).

7.7.1 Construção da Matriz Binária Termo-Documento

Durante a extração do conjunto de termos pelo método de Bruandet (Capítulo 2) supõe-se conhecer o conjunto dos documentos indexados pelos termos. A relação binária representada pela matriz termo-documento é portanto deduzida pela correspondência entre o conjunto dos termos e o conjunto dos documentos.

A Figura 7.13 mostra a representação da matriz binária termo-documento que é o ponto de partida para se gerar o grafo termo-documento, que será visto na seção seguinte.

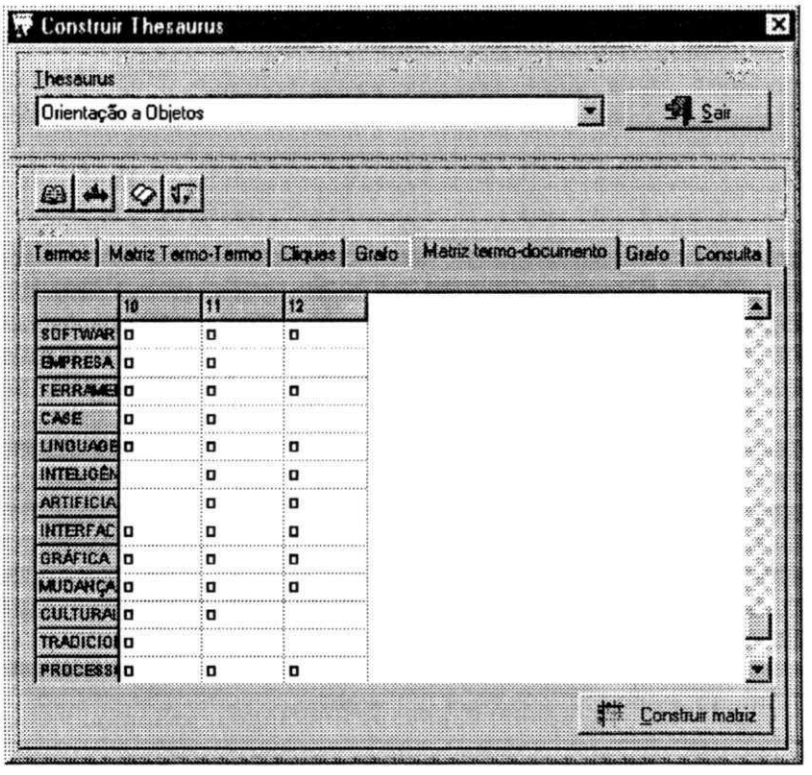


Figura 7.13: Representação da Matriz Binária Termo-Documento

7.7.2 Classificação dos Documentos

Utilizando a *Decomposição Retangular de uma Relação Binária*, organiza-se os diferentes retângulos ótimos extraídos da matriz termo-documento na forma de um grafo (Capítulo 5).

A Figura 7.14 mostra de forma textual o grafo que representa organização hierárquica dos documentos. O grafo de retângulos ótimos que representa a base documental permite uma pesquisa simétrica, seja através dos termos, seja através dos documentos.

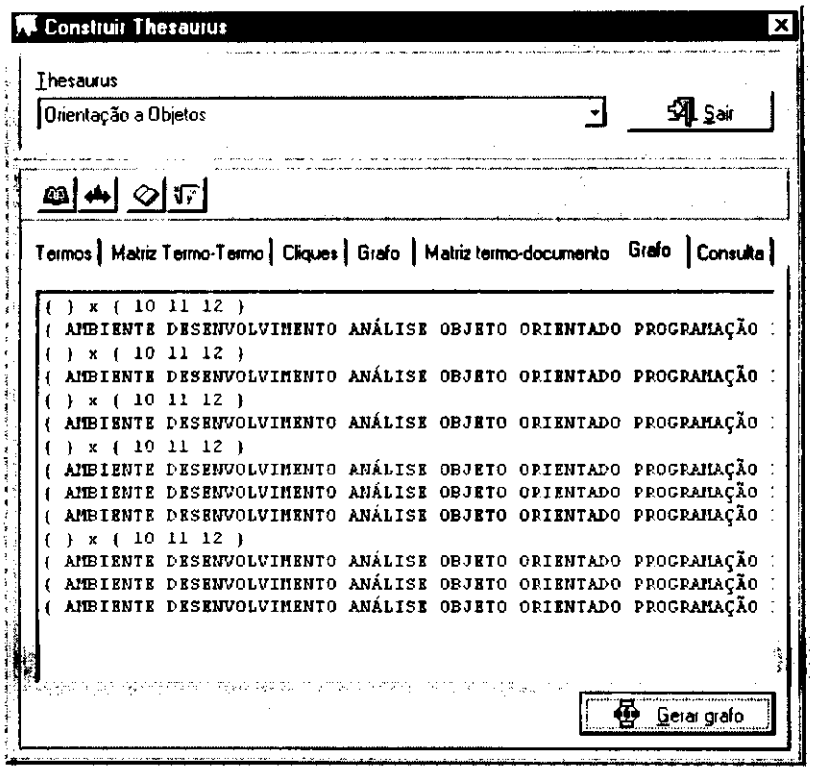


Figura 7.14: Representação textual do grafo termo-documento

7.8 Consultas

No Capítulo 6 são propostas diversas formas de consulta a um Sistema Documental Retangular. Porém, para simplificar da aplicação, optou-se por uma maneira mais simples de consulta, ficando as propostas apresentadas no Capítulo 6 como sugestões para estudos e possível implementações em trabalhos futuros.

A Figura 7.15 mostra a janela do sistema onde são executadas consultas no *Thesaurus* Retangular. No campo **Consulta** o usuário digita um ou mais termos de pesquisa separados por espaço qualquer outra caracter que identifique o final de uma palavra. Após a execução da consulta, o usuário obterá na lista de **Termos**, os termos relacionados à consulta, obtidos através da pesquisa no *thesaurus* (grafo termo-termo). Na lista de **Documentos** serão apresentados os nomes dos documentos (arquivos) indexados pelos termos da consulta, que são obtidos através da pesquisa no grafo termo-documento.

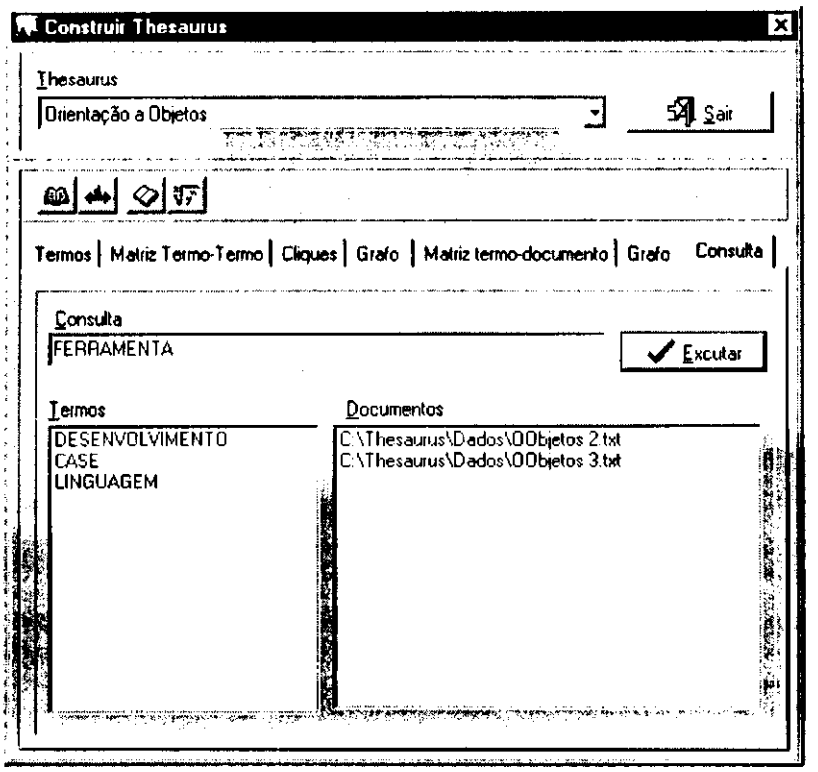


Figura 7.15: Janela para a execução de consultas

7.9 Resultados Experimentais

Durante o desenvolvimento da aplicação diversos testes foram feitos utilizando textos curtos. Após o término completo do Sistema, foi realizado um teste maior utilizando três artigos sobre Orientação a Objetos extraídos da revista *Developers' Magazine* (Ano 1, Nº 4 – Dezembro de 1996). São eles:

- ✓ ***Orientação a Objetos: Mito ou Realidade?*** [Cezar Taurion]
- ✓ ***Orientação a Objetos: o Que Podemos Fazer Hoje e Amanhã?*** [Átila Belloquim]
- ✓ ***Orientação a Objetos: uma Saída Para a Crise do Software?*** [Antonio Carlos Scola]

Os três textos juntos somaram um total de 5.099 palavras e 80 parágrafos.

Do total de palavras existentes, apenas 1017 palavras foram encontradas no dicionário e seleccionadas. Desse total de palavras, após normalização, restou um conjunto de 151 termos distintos.

Utilizando-se o valor **3** como parâmetro **n**, referente a **força de ligação** (Equações 8 e 9, Capítulo 2) e **força mínima de ligação** igual a 0,1 (Seção 2.4, Capítulo 2), foram encontrados os seguintes cliques:

[AMBIENTE DESENVOLVIMENTO]
[ANÁLISE METODOLOGIA]
[ANÁLISE OBJETO ORIENTADO]
[ANÁLISE PROGRAMAÇÃO]
[ANÁLISE PROJETO]
[APLICAÇÃO DESENVOLVIMENTO]
[APLICAÇÃO SISTEMA]
[BANCO DADOS]
[BIBLIOTECA CLASSE]
[CLASSE NÍVEL]
[CLIENTE SERVIDOR]
[CONCEITO ENCAPSULAMENTO]
[CONCEITO HERANÇA]
[DADOS CLIENTE]
[DADOS ESTRUTURA]
[FERRAMENTA CASE]
[FERRAMENTA DESENVOLVIMENTO]
[FERRAMENTA LINGUAGEM]
[HERANÇA POLIMORFISMO]
INTELIGÊNCIA ARTIFICIAL]
[INTERFACE GRÁFICA]
[MÉTODO TRADICIONAL]
[MUDANÇA CULTURAL]
[OBJETO CLASSE]
[OBJETO COMPONENTE]
[OBJETO CONCEITO]
[OBJETO DADOS]
[OBJETO DESENVOLVIMENTO ORIENTAÇÃO]
[OBJETO MÉTODO]
[OBJETO MODELO]
[OBJETO NEGÓCIO]
[OBJETO REAL]
[OBJETO RELACIONAL]
[OBJETO TECNOLOGIA]
[PROCESSO DESENVOLVIMENTO]
[PROGRAMAÇÃO ESTRUTURADA]
[SISTEMA ESTRUTURADO]
[SOFTWARE DESENVOLVIMENTO]
[SOFTWARE EMPRESA]
[SOFTWARE NEGÓCIO]

Após a Decomposição Retangular, os quarenta cliques apresentados acima geraram um grafo contendo 25 nós (retângulos ótimos).

8. Conclusão e Sugestões para Trabalhos Futuros

O trabalho de Bruandet [Bruandet 1989b] fornece a base teórica para a fase inicial da construção do *thesaurus*, que é a extração de termos do conjunto de documentos. A análise dos resultados é facilmente realizada a através dos *cliques*. No Sistema implementado, a maioria dos *cliques* selecionados mostraram-se pertinentes aos assuntos ou idéias contidas nos documentos. Quando o número de *cliques* inexpressivos for excessivo, os dois parâmetros disponíveis permitem "calibrar" a aplicação, melhorando a "qualidade" dos *cliques* ou eliminando os que possuem pouca relevância.

Uma análise dos *cliques* permite precisar o sentido dos termos utilizados através de seu contexto. Por exemplo, se em um mesmo *clique* aparecem os termos *comentário*, *programação* e *linguagem* e, se o usuário estiver familiarizado com a informática, ele poderá deduzir que os documentos só tratam de comentários escritos em programas. Ele sabe, portanto que não poderá obter documentos que tratam de comentários feitos pela imprensa sobre fatos políticos, por exemplo.

O método de *Decomposição Retangular de uma Relação Binária* [Gammoudi 1993] permite a extração de retângulos ótimos de uma relação binária, representada por uma matriz binária. Este método é utilizado não só na construção de um *Thesaurus Retangular* mas também na organização hierárquica dos documentos, fornecendo uma uniformidade

conceitual ao método e permitindo uma grande economia de código para sua implementação. O método permite classificar os termos e os documentos de uma forma automática sem recorrer a ferramentas estatísticas. O grau de generalidade/especificidade e de vizinhança fornece aos usuários ferramentas para reduzir ou aumentar o escopo de suas consultas. A utilização dessas ferramentas dependem fortemente do comportamento do usuário. O estudo do comportamento do usuário se mostra complexo e ultrapassa os limites desse trabalho, podendo ser abordado por trabalhos futuros.

As consultas em um *Thesaurus Retangular* são executadas utilizando os dois reticulados gerados. Através do "reticulado termo-termo", são recuperados os termos relacionados aos termos descritos na consulta do usuário. Através do "reticulado termo-documento" são recuperados os documentos onde estão presentes os termos inquiridos na consulta do usuário.

Como pode ser visto na Figura 7.7 (Capítulo 7), um *thesaurus* é especificado através de um identificador (nome) e por um conjunto de documentos (arquivos). Uma limitação do Sistema é a possibilidade de se utilizar apenas arquivos do tipo texto. Fica como sugestão para futuras implementações, a possibilidade de se trabalhar com os diversos outros tipos de arquivos, gerados pelos diversos editores de textos do mercado.

Durante o processo de extração de termos dos documentos, é utilizado um único dicionário previamente definido. Uma possível melhoria seria a possibilidade de se definir vários dicionários relativos a assuntos específicos. Na definição de um *thesaurus*, o usuário poderia associar um dos dicionários existentes, que seria utilizado durante o processo de extração de termos. Dessa maneira, por exemplo, poderia se definir um *thesaurus* contendo diversos textos da área médica e associa-lo a um dicionário contendo termos médicos importantes.

O Sistema utiliza um dicionário (Figura 7.5) para reduzir cada palavra extraída dos textos a uma forma normalizada (termo) e para recuperar a categoria gramatical desses termos. Caso uma palavra não seja encontrada no dicionário, ela é descartada. As tarefas de normalização e categorização das palavras poderiam ser feitas em uma etapa anterior à utilização do dicionário, através do estudo de prefixos e sufixos.

Uma outra proposta é a sofisticação do dicionário, podendo-se acrescentar novos recursos como o tratamento de homônimos e de contexto, outros idiomas (dicionário multilíngue), etc.

Neste trabalho implementamos as heurísticas propostas em [Gammoudi, 1993]. Visando a diminuição do tempo de decomposição de uma relação binária, futuras implementações poderiam tratar da melhoria das heurísticas, baseando-se sempre nos mesmos conceitos teóricos. Além disso em uma base de dados documental não foi considerado o aspecto dinâmico, isto é, o caso onde a base de dados é modificada devido a uma operação de inserção, alteração ou supressão de um documento. Sendo a semântica da base representada por uma relação binária, uma tal operação implica a modificação desta relação. Nesta ótica, as seguintes questões se fazem:

- ✓ ***É preciso refazer a decomposição da relação ?***
- ✓ ***Pode-se explorar os resultados da decomposição da relação antes de sua atualização ?***

Estas duas questões dizem respeito tanto à construção do *thesaurus* retangular como ao grafo que representa a base documental.

9. Anexos

Anexo A

Algorítmo para extração dos cliques de um grafo

```
S := ∅;
CLIQUE(V, ∅);

CLIQUE(N, D)
inicio
     se N ∪ D = ∅ então
       S é um clique;
     senão
     inicio
       se N ≠ ∅ então
       inicio
         (* explorar primeira sub-árvore *)
         f := vértice em N;
         EXPLORE( f );
         (* explorar o restante da sub-árvore *)
         enquanto N ∩ (V - Adjacentes( f )) ≠ ∅ faça
         inicio
           v := vértice em N ∩ (V - Adjacentes( f ));
           EXPLORE( v );
         fim;
       fim;
     fim
fim;

EXPLORE(u)
inicio
     N := N - {u};
     S := S ∪ {u};
     CLIQUE(N ∩ Adjacentes(u), D ∩ Adjacentes(u));
     S := S - {u};
     D := D ∪ {u};
fim;
```

Anexo B

Algoritmo para pesquisa de um retângulo ótimo contendo um elemento

```
Retangulo_Otimo(R; (a, b); RE);
     (* R é uma relação binária,
       (a, b) é um elemento de R,
       RE é a variável destinada a conter um retângulo ótimo de R,
       contendo o elemento (a, b) *)
inicio
     ParesSelecionados := (a, b);
     PR=Φ_R(a, b);
     X:=a; Y:=b;
     ω(vazio) := -2; (por convenção)
     enquanto PR' não for um retângulo faça
     inicio
       PR0:=∅; PR1:=PR - ParesSelecionados;
       enquanto restar elementos em PR1 contendo X como argumento
                 ou Y como imagem faça
       inicio
         escolher um elemento (x, y) de PR1 da forma (X, y) ou (x, Y);
         PR1:=PR1 - {(x, y)};
         PR':=Φ_PR(X, Y);
         se PR'=PR então   (* a relação elementar PR' é ignorada *)
           ParesSelecionados:=ParesSelecionados ∪ {(X, Y)};
         senão
         inicio
           se ω(PR') > ω(PR0) então
           inicio
             PR0:=PR';
             X':=x;
             Y':=y;
           fim;
         fim;
       fim;
       X:=X'; Y:=Y';
       ParesSelecionados ∪ {(X, Y)};
       PR':=PR0;
     fim;
     RE:=PR0;
fim;
```

Anexo C

Algoritmo para seleção de uma Cobertura Mínima

```
Cobertura
inicio
     R0 := R;
     para todo par(a, b) de R faça
     inicio
        PR':= Φ_R(a, b);
        (* Φ_R(a, b) é uma relação elementar contendo o par (a, b) *)
        ω(PR'):=(r/(d * c)) * (r - (d + c ) );
        (*r=cardinal(PR'); d=cardinal(dom(PR')); c=cardinal(cod(PR'))*)
     fim;
     ordenar os elementos (a, b) de R em ordem decrescente de ω(PR');
     enquanto R <> ∅ faça
     inicio
        Recebe o par (a, b) de R que maximiza o ganho
        Retangulo_Otimo(R0, (a, b), PR');
        (* eliminação de certos elementos que se repetem *)
        PR':=PR' ∩ R;
        PR':=PR'^++ ; (* PR'^++ é o fechamento retangular de PR' *)
        R:=R - PR';
     fim;
fim;
```

Anexo D

Algoritmo de construção de um Grafo de Retângulos Ótimos

```
Grafo_de_Retangulos(LR)
(* LR é uma lista contendo os retângulos *)
inicio
     Nos := (Ø, E);
     LNS := Ø;
     Nivel:=1;
     enquanto Nivel < n faça
       enquanto LR(Nivel) <> Ø faça
       inicio
         RecebeElemento(RE, LR(Nivel));
         para todo RE_i em LR(Nivel-1) faça
           se RE_i está em RE então
           inicio
             atribui(RE_i, RE);
             pred := verdadeiro;
             RE_i.marca := verdadeiro;
           fim;
         se pred=falso entao
           para todo RE_i em LNS faça
                se RE_i está em RE então
                inicio
                  atribui(RE_i, RE);
                  pred := verdadeiro;
                  RE_i.marca := verdadeiro;
                  suprimir(RE, LNS);
                fim
         se pred = falso então
           atribui(Nos, RE);
       fim
       para todo RE_i em LR(nivel-1) faça
         se RE_i.marca = falso então
           Inserir(RE_i, LNS);
       nivel := nivel + 1;
     fim
     para todo RE em LNS faça
       atribui(Nos, RE);
fim;
```

Referências

ABE, JAIR MINORO & PAPAVERO, NELSON. 1992. *Teoria Intuitiva dos Conjuntos*. Makron Books.

ATTAR, R. & A. S. FRAENKEL. 1977. *Local Feedback in Full-Text Retrieval System*. Jornal of the ACM, Vol. 24, N° 3, Julho, 397-417.

BELKHITER, N., DESHARNAIS, J., ENNIS, G., GAMMOUDI, M.M., JAOUA, A., MOUKAM, T., LE THANH, N., REGUIG, M. 1992. *Propriétés formelles des relations rectangulaires: Application à l'organisation et à l'interrogation d'une base de donnés documentaire*. Université Laval, Département d'informatique, Rapport de Recherche, DIUL-RR-9204.

BRUANDET, MARIE-FRANCE. 1980a. *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, MARIE-FRANCE. 1980b. *A Propos de la Construction Automatique d'un Thésaurus dans un Système de Recherche d'Information (Systeme Documentaire)*. IMAG. Rapport de Recherche (Relatório Récnico) n° 229. Novembro, 1980.

BRUANDET, MARIE-FRANCE. 1981. *Notion de Concept pour la Construction Automatique d'un Thesaurus Evolutif*. AFCET Informatique. Actes du congrès de l'Afcet. 18-20 de Novembro, 1981. Editions Hommes et Techniques.

BRUANDET, MARIE-FRANCE. 1982a. *Concept Notion for Automatic and Dynamic Thesaurus Updating.* Conference Proceedings International Conference on System Documentation - SIGDOC. Carson, California. 23-23 de Janeiro, 1982.

BRUANDET, M-F., CHIARAMELLA, Y., KERKOUBA, D., 1982b. *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, MARIE-FRANCE. 1985. *Modele Partiel de Connaissances pour un Systeme de Recherche d'Informations.* Recherche d'Informations Assistée par Ordinateur - RIAO'85. Grenoble, France. 18-20 de Março, 1985.

BRUANDET, MARIE-FRANCE. 1989a. *Outline of a Knowledge-Base Model for an Intelligent Information Retrieval System.* Information Processing & Management. Vol. 25, N° 1, pp. 89-115. 1989.

BRUANDET, MARIE-FRANCE. 1989b. *Construction Automatique d'une Base de Connaissances du Domaine dans un Systeme de Recherche d'Informations.* Document fourni pour la soutenance du Diplôme d'Habilitation à Diriger des Recherches de L'Université Joseph Fourier de Grenoble. 13 de Março, 1989.

EVERETT, C.J. 1944. *Closure Operators and Galois Theory in Lattices.* Trans. Amer. Math. Soc., (55). Pp. 514-525.

GAMMOUDI, MOHAMED MOHSEN. 1993. *Méthode de Décomposition Rectangulaire d'une Relation Binaire: Une base formelle et uniforme pour la génération automatique des thesaurus et la recherche documentaire.* Tese de doutorado. Universite de Nice — Sophia Antipolis Ecole Doctorale des Sciences pour L'Ingenieur.

GUENOCHE, A., MONJARDET, B. 1987. *Méthodes ordinales et conbinatoires en analyse des données.* Math. Sci. Hum., (100): 893-898.

HENDERSON, K. 1996. *Database Developer's Guide with Delphi 2.* SAMS Plublishing. Borland Press.

LÉVY, PIERRE. 1993. *As Tecnologias da Inteligência - O Futuro do Pensamento na Era da Informática..* Editora 34.

MARTIN, JAMES. 1992. *Hiperdocumentos e como criá-los.* Editora Campus.

PACHECO, XAVIER & TEIXEIRA, STEVE. 1996. *Delphi 2 Developer's Guide.* SAMS Plublishing. Borland Press.

REINGOLD, M.E. NIEVERGELT, J. DEO, N. 1977. *Combinatorial Algorithms • Theory and Practice*. Prentice Hall, Englewood Cliffs, NJ, 353-359..

WILLE, R. 1985. *Finite Distributive Lattice as Concept Lattices*. Atti Inc. Logica Matematica, (2). Pp635-648.

Bibliografia

BARBER, A.S., BARRACLOUGH, E.D. AND GRAY, W.A. 1973. *On-line information retrieval as a scientist's tool.* Information Storage and Retrieval, 9, 429-44.

BARBOUT, M., MANJARDET, B. 1970. *Order et classification, algèbre et conbinatoire (Tome 2).* Hachette, Paris.

BAR-HILLEI, U. 1964. *Language and Information. 'Selected Essays on their Theory and Application.* Addison-Wesley, Reading, Massachusetts.

BAPTISTE, P., FAVREL, J. 1984. *Résolution de problèmes d'ordonnancement par les trellis de Galois et les graphes d'intervalle.* R.A.I.R.O. Automatique/Systems Analysis and Control, 18 (4): 405-416.

CHIARAMELLA, Y., KERKOUBA, D., BRUANDET, M.F., 1986. *Integration d'une Fonction Documentaire dans un Atelier de Logiciel.* CONCERTO - Atelier de Logiciel. 4-6 Fevereiro, 1986.

CHIARAMELLA, Y., B. DEFUDE, BRUANDET, M.F., KERKOUBA, D., 1986. *IOTA: A Full Text Information Retrieval System.* ACM Comference on Research and Development in Information Retrieval. Pisa - Italy. 8-10 Setembro, 1986.

CLEVERDON, C.W., MILLS, J. E KEEN, M. 1966. *Factors Determining the Performance of Indexing Systems*, Vol. I, *Design*, Vol II, *Test Results*, ASLIB Cranfield Project, Cranfield.

DOYLE, L.B. 1965. *Is automatic classification a reasonable apllication of statistical analysis of text?* Jornal of the ACM, 12, 473-489.

DAVEY, B.A., PRIESTLEY, H.A. 1990. *Introduction to Lattices and Order.* Cambridge University Press, pp. 222-240.

DELOBEL, C. 1978. *Normalization and Hierarchical Depencencies in the Relational Data Model. ACM Trans. on Database Systems,* 2(3): 201-222.

FRAKES, W. B., BAEZA-YATES, R. (eds), 1992. *Information Retrieval: Data Strutures & Algorithms,* Prentice Hall.

GODIN, R. MISSAOUI, R., APRIL, A. 1992. *Experimental Comparison of Navigation in a Galois Lattice with Conventional Information Retrieval Methods.* International Journal of Man-Machine Studies.

GOOD, I.J. 1958. *Speculations concerning information retrieval,* Research Report PC-78, IBM Research Centre, Yorktown Heights, New York.

GREFENSTETTE, GREGORY. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers.

LANCASTER, F.W. 1991. *Indexação e Resumo: teoria e prática.* Briquet de Lemos/Livros.

LUHN, H.P. 1957. *A statistical approach to mechanised encoding and searching of library information.* IBM Journal of Research and Development, **1**, 309-317.

LUHN, H.P. 1958. *The automátic creation of literature abstracts,* IBM Journal of Research and Development, **2**, 159-165.

MCCARN, D.B. & LEITER, J. 1973. *On-line services in medicine and beyond,* Science, **181**, 318-324.

MARON, M.E. & KUHNS, J.L. 1960. *On relevance, probabilistic indexing and information retrieval.* Journal of the ACM. 7, 216-244.

SALTON, GERARD & M. E. LESK. 1965. *The SMART Automatic Document System — An Illustration.* Communications of the ACM, Vol. 8, N° 6, Junho.

SALTON, GERARD. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Printice-Hall

SALTON, GERARD. 1975. *Dynamic Information and Library Processing.* Prentice-Hall.

SALTON, GERARD. 1989. *Automatic Text Processing*. Addison Wesley.

SPARCK JONES, K. 1971. *Automatic Keyword Classification for Information Retrieval*, Butterworths, London.

SENKO, M.E. 1969. *Information storage and retrieval system. Advances in Information Systems Science*, Plenum Press, New York.

WESSEL, ANDEW E. 1975. *Computer-Aided Information Retrieval*. Melville Publishing Company. Los Angeles, California.

Ciência

Começo a ver no escuro
um novo tom
de escuro.

Começo a ver o visto
e me incluo
no muro.

Começo a distinguir
um sonilho, se tanto,
de ruga.

E a esmerilhar a graça
da vida, em sua
fuga.

Carlos Drummond de Andrade

The time is gone the song is over, thought I'd something more to say

Pink Floyd