Termos, relacionamentos e representatividade na indexação de texto para recuperação de informação

Marco Gonzalez*,**
PUCRS e UFRGS

Vera L. S de Lima*
José V. de Lima**
UFRGS

Resumo: Uma das fases da recuperação de informação é a indexação dos textos dos documentos. Nesta fase, um conjunto de descritores (termos e/ou relacionamentos entre termos) descreve conceitos (atômicos e/ou complexos) presentes nos textos. Diversas estratégias com tais finalidades são encontrados na bibliografia, algumas consideram dependência de termos e outras não.
Com o objetivo de apresentar uma visão geral das estratégias de representação de textos que consideram dependência de termos, são descritas quatro experiências onde as representatividades dos relacionamentos dependem dos termos componentes (estratégias com índices múltiplos, com árvore binária, com triplas e com famílias morfológicas), três onde as representatividades dos relacionamentos dependem de suas próprias freqüências de ocorrência (estratégias com expressões de índice, com pares lematizados e com expressões ternárias), duas onde os relacionamentos são reconhecidos mas não são utilizados como descritores (estratégias com nodos temáticos e com conexões gramaticais) e uma experiência onde os relacionamentos são eminentemente estatísticos (estratégia com bitermos).

* Faculdade de Informática. {gonzalez, vera} @inf.pucrs.br
** Instituto de Informática. valdeni@inf.ufrgs.br

1 Introdução

Indexação de textos é o processo da recuperação de informação (RI) que estabelece descritores (unidades de indexação) dos conteúdos dos textos de uma coleção de documentos, com propósito de busca e classificação dos mesmos para atender consultas em sistemas de RI. Descritores podem descrever conceitos atômicos, sendo termos, ou conceitos complexos, sendo relacionamentos. Cada descritor pode ser mais ou menos representativo do texto de um documento e essa variação deve ser conhecida.

Para a indexação, o texto dos documentos deve ser pré-processado de acordo com a estratégia adotada. São usuais os procedimentos de *toquenização* (identificação de cada item lexical do texto, tal como palavra e pontuação), seleção de descritores e conflação [BAE 99]. Se a estratégia adotada necessita de informação lingüística, outros métodos devem ser considerados. A análise morfo-sintática do texto pode ser obtida através de um *parser* [MAR 03]. Podem ser incluídas, também, anotações para resolução de co-referências [VIE 02]. Essas informações são, geralmente, inseridas através da etiquetagem de texto após a *tokenização*. Um etiquetador gramatical (*part-of-speech tagger*) identifica, com a colocação de uma etiqueta (*tag*), a categoria gramatical de cada palavra do texto. Geralmente é morfológico (identifica somente a categoria morfológica) ou morfo-sintático (identifica também as funções sintáticas).

Após o pré-processamento do texto dos documentos, a geração do espaço de descritores, propriamente dita, acontece. Nesse processo se dá a seleção e a normalização (se houver) dos descritores e, ainda, a obtenção das informações necessárias (geralmente a freqüência de ocorrência) para o cálculo de seus pesos. Tal cálculo avalia a representatividade dos descritores em relação ao conteúdo do texto e à relevância dos conceitos descritos em cada documento. O peso de um descritor é, entretanto, afetado, indiretamente, pelas estratégias adotadas para sua seleção e normalização.

2 Termos

A expressão 'termo' é entendida aqui como uma unidade lexical formada por uma única palavra ou por mais de uma. Neste sentido, a expressão 'termo composto' tem sido usada para indicar uma unidade lexical complexa com características de uma única unidade sintática e semântica [DIA 00], descrevendo um único conceito [STR 96]. Por exemplo, 'boca da noite' será um termo, se

for identificado como uma única unidade lexical, ou, ao contrário, constituirá mais de um termo. Neste último caso, conforme a abordagem, poderá ser reconhecida ou não uma dependência entre os termos 'boca' e 'noite'. Termos (compostos ou não) são, de qualquer forma, descritores.

Salton e MacGill utilizam o gráfico de Luhn [SAL 83] para definir os bons descritores e discutir como obtê-los. Segundo este gráfico, a expressividade dos termos com freqüência de ocorrência intermediária é maior que a dos termos muito ou pouco freqüentes. Bons descritores deveriam ser, então, derivados por transformação dos termos ruins através da construção de frases, ou por transformação dos termos pobres em termos bons através de relações encontradas em um thesaurus. Thesauri são dicionários que não definem as palavras, como tradicionalmente é feito, mas as organizam como conceitos relacionados semanticamente entre si.

A usual eliminação de s*topwords* faz parte do processo de seleção de descritores. *Stopwords* são palavras como preposições, artigos e conjunções, que podem ser descartadas em fase de indexação e na consulta. Essa eliminação, com conseqüente redução do espaço de descritores, justifica-se porque as *stopwords*, tidas como descritores ruins, são consideradas desnecessárias por terem pouca representatividade. Ao serem descartadas as preposições, por exemplo, ganha-se em economia mas perde-se em representatividade. Por exemplo: 'caixa de vidro' e 'caixa para vidro' têm, evidentemente, significados diferentes e esta diferença não pode ser representada, neste caso, sem as preposições.

Processos para normalização lexical (ver próxima Seção) também geram descritores com maior expressividade, pois as freqüências das formas normalizadas serão maiores que as dos termos originais, mais raros por suas variações morfológicas. Por outro lado, há descritores originalmente pobres que não conseguem aumentar sua expressividade, mesmo através da normalização, levando em conta apenas a freqüência de ocorrência. Entretanto, são bons representantes de conceitos relevantes presentes no texto. Neles o autor pode estar centralizando uma idéia importante que pretende comunicar, ainda que não concretize isso, essencialmente, através da freqüência com que os usa. Por exemplo, um substantivo com a função de núcleo do sujeito de uma oração terá maior evidência que se utilizado como adjunto adnominal. Tal evidência, portanto, não depende apenas da freqüência de ocorrência.

2.1 *Normalização lingüística*

A seleção dos descritores, a quantidade dos mesmos e o peso de cada um podem ser afetados pela estratégia adotada para normalização lingüística. O reconhecimento de variações lingüísticas encontradas em um texto permite o controle de vocabulário [JAC 97]. Tal controle determina termos preferenciais a serem usados como descritores, ou seja, os seleciona, os restringe em número e, em conseqüência, influencia o cálculo da representatividade dos mesmos.

Há três tipos de normalização lingüística [ARA 00, SAV 03]: sintática, léxico-semântica e morfológica (ver Tabela 1).

Tabela 1. Normalização lingüística

normalização		método usual
lexical	morfológica	conflação
	léxico-semântica	busca de relações semânticas em thesaurus
sintática		aplicação de regras gramaticais

A normalização sintática ocorre quando há a transformação de frases semanticamente equivalentes mas sintaticamente diferentes, em uma forma única e representativa das mesmas, como 'eficiente processo rápido' e 'processo rápido e eficiente', que poderiam ter uma representação comum.

A normalização léxico-semântica ocorre quando são utilizados relacionamentos semânticos (como a sinonímia) para substituir palavras morfologicamente distintas por uma única forma que representa o conceito referenciado.

Em nível lexical há dois extremos de normalização. De um lado há a normalização léxico-semântica, através de busca de sinônimos, por exemplo, em thesaurus, considerando informações terminológicas [JAC 99]. Em outro extremo está a normalização morfológica, através da conflação, que explora similaridades morfológicas inferindo proximidades conceituais.

A normalização morfológica ocorre quando há redução das formas flexionais de uma palavra, através de conflação, a uma forma única que procura representar um conceito ou uma classe de conceitos. Os processos mais comuns de conflação são o *stemming* [FRA 92, KRO 93, ALL 03] e a lematização [ARA 00, KOR 04]. S*temming* é um processo que reduz ao mesmo *stem* (parte fundamental semelhante ao radical) palavras que se diferenciam basicamente pela flexão, como:

stemming(livro) = *stemming*(livros) = livr ou
stemming(caminhada) = *stemming*(caminhei) = caminh.

A lematização reduz as palavras variáveis à correspondente forma canônica: verbos no infinitivo e palavras, como substantivos e adjetivos, no singular e, quando existir, masculino. São exemplos:

lematização(livro) = lematização(livrinho) = livro,
lematização(livre) = lematização(livres) = livre ou
lematização(caminhar) = lematização(caminhei) = caminhar.

A principal diferença entre os resultados de *stemming* e de lematização é que, no primeiro caso, palavras de diferentes categorias morfológicas podem ter o mesmo *stem*, como:

stemming(construiu) = *stemming*(construções) = constru,

enquanto que, na lematização, a categoria morfológica é mantida:

lematização(construiu) = construir ≠ lematização(construções) = construção.

Conforme Braschleer e Ripplinger [BRA 99], os benefícios do *stemming* são reconhecidos na RI. Este processo (i) reduz o número de descritores e o tamanho do arquivo de índice, e (ii) torna a recuperação independente da forma com que o termo ocorre na consulta. O mesmo pode ser dito sobre a lematização. Isso confirma que os processos de conflação têm um segundo efeito, além da normalização lexical em si: a economia na quantidade de descritores e no espaço de memória necessário para armazená-los.

Apesar dos benefícios desses métodos de conflação, alguns problemas ainda estão por ser resolvidos. O *stemming*, por exemplo, não tem sucesso com termos onde a flexão é raramente usada ou inexiste (por exemplo, nomes próprios) [BRA 99]. A análise flexional e morfológica de termos compostos também é problemática, mesmo para a língua Inglesa [SAV 03]. Uma solução, nesses casos, é a decomposição do termo e a aplicação da normalização, separadamente, a cada componente, especialmente na lematização [KOR 04].

Alguns significados podem ser perdidos no *stemming* e palavras de famílias de significados diferentes podem ser agrupadas. Esses erros podem ser difíceis de detectar e corrigir em sistemas automáticos, requerendo esforço adicional para tratamento de exceções. Seriam os casos de:

stemming(livro) = *stemming*(livre) = livr e
stemming(caminhada) = *stemming*(caminhão) = caminh.

Enquanto um algoritmo simples de *stemming* é suficiente para Inglês, estratégias mais sofisticadas são necessárias para idiomas com morfologia flexional complexa, como Alemão, Espanhol, Finlandês, Francês e Português. *Stemmers* para tais idiomas apresentam elevado custo computacional [VIL 02].

Experimentos com Espanhol [VIL 02] e Finlandês [KOR 04] concluem que a lematização produz, na RI, melhores resultados que o *stemming*. Entretanto, a lematização também apresenta problemas. Algumas palavras pertencentes à mesma família de significados podem não ser normalizadas, como:

lematização(livre) = livre ≠ lematização(liberdade) = liberdade

e

lematização(caminhei) = caminhar ≠ lematização(caminhada) = caminhada.

3 Cálculo da representatividade

Após a seleção e a normalização dos descritores, é necessário calcular a representatividade, que é uma das propriedades básicas de um descritor. O grau de representatividade de um descritor i em um documento d é dado pelo peso $W_{i,d}$. Na abordagem vetorial, um dos modos usuais de calcular esse peso é o seguinte [SAL 88]:

$$W_{i,d} = \frac{w_{i,d} IDF_i}{\sqrt{\sum_j \left(w_{j,d} IDF_j\right)^2}} \quad (1)$$

onde: j representa cada um dos descritores do espaço de descritores; $w_{j,d} = f_{j,d}$ ou $w_{j,d} = \frac{f_{j,d}}{\max f_d}$ (o mesmo vale para $w_{i,d}$); $f_{j,d}$ = freqüência de ocorrência de j em d; max f_d = máxima freqüência de ocorrência dos descritores de d; $IDFj = \log \frac{N}{df_j}$ (o mesmo vale para IDF_i); N = número de documentos na coleção; e df_j = número de documentos onde j ocorre.

No modelo probabilístico, um dos esquemas usuais de cálculo do peso de um descritor i em um documento d corresponde à fórmula Okapi BM25 [ROB 94, SPA 00]:

$$W_{i,d} = \frac{w_{i,d}(k_1 + 1)}{k_1((1-b) + b\frac{DL_d}{AVDL}) + w_{i,d}} IDF_i \quad (2)$$

onde: $w_{i,d} = f_{i,d}$ é a freqüência de ocorrência de i em d; k_1 e b são parâmetros (discutidos adiante); DL_d é o comprimento (quantidade de palavras) do documento d; $AVDL$ é o comprimento médio dos documentos da coleção; e IDF_i é o mesmo fator utilizado na Equação 1.

O fator *IDF* (*inverse document frequency*) é utilizado para penalizar descritores que ocorrem com muita freqüência na coleção [HIE 00]. O uso da freqüência de ocorrência e do fator *IDF* caracteriza esquemas conhecidos como TF.IDF [SAL 88, HIE 00].

O parâmetro k_1 (da Equação 2) é utilizado para correção da freqüência. Com valores de k_1 entre 1,2 e 2 (valores usuais – especialmente 1,2), ocorrências adicionais do descritor, acima da terceira ou quarta ocorrência em um documento, têm impacto mínimo no cálculo [SPA 00].

A utilização do parâmetro b (da Equação 2) e do comprimento (quantidade de palavras) do documento está relacionada às hipóteses do escopo e da verbosidade [ROB 94]. Na hipótese do escopo, documentos mais longos têm mais informação que documentos menos longos. Na hipótese da verbosidade, documentos mais longos possuem escopo similar ao de um documento menos longo, simplesmente usam mais palavras. Na prática, documentos de coleções reais combinam estes dois efeitos [ROB 94]. Com relação ao parâmetro b da Equação 2, se $b = 1$, anula-se a hipótese do escopo predominando somente a da verbosidade; se b assumir valores menores que 1, diminui a importância da verbosidade; e se $b = 0$, a hipótese da verbosidade se anula. Valores usuais ficam em torno de 0,75 [SPA 00].

4 Relacionamentos

Até aqui, foram discutidos a seleção, a normalização e o cálculo da representatividade sempre tendo em vista descritores constituídos como termos. Entretanto, um espaço de descritores, conforme a abordagem adotada, pode incluir também os relacionamentos entre esses termos.

Alguns sistemas de RI permitem consultas através de frases, como ‘defesa eficiente’ e ‘feira de domingo’. Nestes casos, o usuário não estaria interessado em qualquer ‘defesa’ ou em qualquer ‘feira’. A coincidência exata poderia garantir, assim, o atendi-

mento da consulta. Entretanto, nem sempre a coincidência exata é satisfatória. Para os exemplos citados, a expressão 'defender eficientemente' poderia ser relevante para a consulta 'defesa eficiente'. Da mesma forma, 'feira dominical' poderia interessar ao usuário da consulta 'feira de domingo'. Isto ocorre porque construções sintaticamente diferentes podem ter o mesmo significado.

Como representar esses significados através de relacionamentos? Por exemplo, o trecho '... têm preocupado os pesquisadores' pode ser representado através de um dos seguintes tipos de relacionamentos: o par modificado-modificador 'pesquisador-preocupado', o bigrama '(preocupado,pesquisador)', o sintagma nominal 'pesquisador preocupado', ou algum outro formato, como a expressão ternária 'preocupação-de-pesquisador' e a relação binária 'de(preocupação,pesquisador)'. Dois, dentre esses tipos de relacionamentos, têm características especiais: os bigramas e os sintagmas nominais.

Há duas diferenças importantes entre a descrição produzida pelos bigramas e a de outros tipos de relacionamentos. A primeira diferença é a inviável descrição de conceito pretendida através de alguns bigramas, como '(ferro,sopa)', capturado de 'panela de ferro com sopa', por exemplo. A segunda diferença é que, quando a descrição é viável, bigramas como '(a,x)' e '(b,x)' representam acepções diferentes do descritor 'x': uma, quando antecedido por 'a', outra, por 'b', cada uma com sua representatividade. Por exemplo, nos bigramas '(sentar,banco)' e '(depositar,banco)', o termo 'banco' produz descrições diferentes em cada um deles, sendo os termos 'sentar' e 'depositar' (importantes) coadjuvantes.

Sintagmas nominais constituem outro caso especial de tipo de relacionamento porque podem representar tanto conceitos complexos quanto atômicos. Conceitos atômicos podem ser descritos por sintagmas nominais formados por um único substantivo, como 'noite', ou mesmo por mais de uma palavra, como 'boca da noite'. O sintagma nominal 'boca da noite' poderá ser um termo (composto), conforme já foi mencionado, ou um relacionamento. De qualquer forma será um descritor.

Tendo sido estabelecidos os formatos para representar dependências de termos, como identificá-las? Diversos pesquisadores têm analisado a aplicação de conhecimentos estatísticos e lingüísticos no tratamento desse problema. Fagan [FAG 87] analisou a representação de relacionamentos estatísticos e sintáticos e apontou vantagens para a segunda abordagem, como a capacidade de identificar relacionamentos entre palavras não adjacentes, como entre as palavras 'algoritmos' e 'concorrentes' em 'algoritmos seqüenciais e concorrentes', por exemplo.

Há ainda outras questões em aberto, além da identificação das dependências entre termos. Por exemplo, é preciso decidir como a representatividade dos relacionamentos deve ser calculada e como esses descritores devem ser normalizados. O encaminhamento dessas questões passa pela especificação completa do espaço de descritores.

5 Exemplos de estratégias com dependência de termos

A seguir são, então, descritos alguns trabalhos selecionados que exemplificam alternativas para representação de textos visando a indexação e, ao final, é apresentada uma análise comparativa dos mesmos.

5.1 Expressões de índice

Wondergem, Bruza e co-autores [BRU 91, WON 00, WON 00a] descrevem o que chamam de 'expressões de índice'. Tais expressões podem caracterizar o conteúdo de um texto através de construções denominadas 'lithoids'. As expressões de índice apresentam o seguinte padrão, descrito utilizando-se formalismo EBNF:

termo { *conector termo* }*

Essas expressões estão relacionadas a sintagmas nominais, sendo que: (i) um termo pode ser uma palavra-chave, um conceito, uma denotação ou um valor de atributo; e (ii) um conector, que representa uma relação entre termos, pode ser uma preposição, um verbo no gerúndio ou o conector nulo, representado por '**.**' (ponto). Alguns tipos de relacionamentos, representados principalmente através de preposições, são: posse ou ação-objeto (*of*), ação-agente (*by*), posição (*in*, *on*, ...), associação direta (*to*, *on*, *for* e *in*), associação (*with* e *and*) e equivalência (*as*).

A derivação de expressões de índice a partir de um texto é realizada através da construção de árvores de representação de sintagmas nominais como, por exemplo, 'a poluição da água por metais' (ver Figura 1). Após a remoção de *stopwords* (com exceção das preposições), os termos remanescentes são sucessivamente processados. No caso de ser encontrado um conector, ele servirá de guia para decidir se a árvore corrente será aprofundada ou alargada.

Um 'lithoid' (Figura 2 – [WON 00a]) pode ser, então, construído a partir da árvore de representação (Figura 1 – adaptada de [BRU 91]). Um 'lithoid', denominado assim em razão de sua estrutura semelhante à de um cristal, é um grafo cujos nodos são constituídos pela expressão de índice inicial e por todas as suas subexpressões. Os arcos conectam cada subexpressão com sua expressão correspondente. O símbolo ε representa uma expressão de índice vazia, que é subexpressão válida para qualquer expressão.

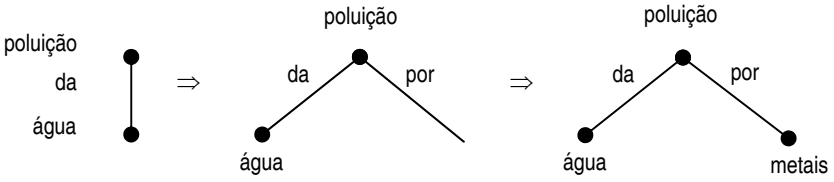

Figura 1. Exemplo de derivação de uma expressão de índice

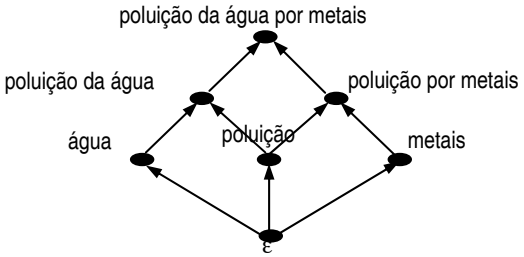

Figura 2. Exemplo de 'lithoid'

Os nodos de um 'lithoid' são vinculados ao texto onde a correspondente expressão ocorre. Assim, alguns nodos podem estar vinculados e outros não. No caso da Figura 2, se os nodos 'poluição por metais', 'poluição' e 'metais' estiverem vinculados a um determinado texto, isto significaria que ele trata de temas associados a estas expressões. O mesmo texto pode não estar vinculado ao nodo 'água', por exemplo.

Strzalkowski e co-autores [STR 96, STR 99] propõem um sistema onde o texto dos documentos é inicialmente processado através de um analisador sintático. Uma gramática com 400 regras é aplicada sentença a sentença. Relacionamentos são extraídos da árvore de análise sintática construída. Eles são usados como descritores juntamente com os termos. Nessa extração, é considerada a distribuição estatística dos componentes para decidir se a associação entre dois termos é sintaticamente válida e semanticamente significante.

Stopwords (preposições, conjunções, artigos, etc.) são removidas do texto. O *stemmer* morfológico tradicional é substituído por um eliminador de sufixos assistido por dicionário que (i) reduz variantes de palavras a suas respectivas raízes, conforme dicionário, e (ii) converte substantivos derivados de formas verbais (tais como os substantivos, em Inglês, 'implementation' e 'storage') nas raízes dos verbos correspondentes (respectivamente 'implement' e 'store'), também com o auxílio de dicionário.

Nomes próprios são identificados e representados como termos. Os relacionamentos extraídos da árvore de análise sintática são pares modificado-modificador. Esses pares são normalizados sintaticamente. Por exemplo (em Inglês): formas como 'weapon profileration', 'proliferation of weapons' e 'proliferate weapons' são reduzidas à forma 'weapon+proliferate'. Termos compostos (como 'joint venture') podem ser extraídos, no lugar de relacionamentos, de acordo com a distribuição estatística dos componentes de cada par analisado.

São gerados, ao final, quatro arquivos de índices: de termos, de termos compostos, de nomes próprios e de relacionamentos. É concebido um fluxo de busca onde descritores são comparados com a consulta em quatro etapas. Cada um dos arquivos de índices é acessado em cada etapa. Após, os resultados parciais são combinados para estabelecer a classificação dos documentos.

É utilizado o esquema TF.IDF para todos os tipos de descritores. Strzalkowski e co-autores acreditam que este esquema não é apropriado para descritores de diferentes tipos (como termos simples, nomes próprios e relacionamentos). Eles suspeitam que todos os esquemas de cálculo de pesos baseados em freqüência de ocorrência apresentem esta desvantagem. Por esta razão, são introduzidos acréscimos aos pesos dos relacionamentos, através de parâmetros que representam fatores de multiplicação.

5.3 *Nodos temáticos*

Loukachevitch, Dobrov e co-autores [DOB 98, LOU 99, LOU 00] propõem a representação temática de textos através de estrutura hierárquica de nodos temáticos (ou conceituais), para RI e para sumarização automática. A abordagem pressupõe que aquelas cadeias lexicais que caracterizam o tema principal de um texto, normalmente, possuem elementos que ocorrem juntos nas sentenças com maior freqüência que outros elementos de outras cadeias lexicais. É assumido o seguinte conjunto de premissas:

- a coesão de um texto pode ser obtida através de referências, elipses, conjunções e termos semanticamente relacionados;
- a coesão lexical é o tipo mais freqüente de coesão textual, e pode ocorrer através de repetições, de relacionamentos como sinonímia e hiponímia, ou de relações sintagmáticas;
- cadeias lexicais (seqüências de termos conectados através de coesão lexical) estão extremamente relacionadas à estrutura temática do texto, sendo, portanto, de importância crucial para o processamento e para a representação do conteúdo tratado;
- a coesão lexical não está baseada em um conjunto isolado de cadeias lexicais mas em uma rede complexa constituída por relações diversas entre os termos; e
- o tema de um texto pode usualmente ser descrito através de temas menos gerais que, por sua vez, podem ser caracterizados por temas ainda mais específicos, e assim por diante, ou seja, a representação temática de um texto é uma estrutura hierárquica de termos.

Uma representação temática, como é concebida nesta abordagem, é constituída por nodos temáticos. Cada nodo temático possui um termo selecionado como 'centro temático' e outros termos (subtemas) semântica e tematicamente associados a este. A representação é construída através dos seguintes passos: identificação dos termos no texto, resolução de ambigüidades (através de um thesaurus de domínio), construção dos nodos temáticos e determinação do *status* dos nodos temáticos.

São considerados 'centros temáticos' os descritores contidos no título e na primeira sentença do texto, bem como aqueles com elevada freqüência de ocorrência. Fazem parte de um mesmo nodo temático os descritores relacionados aos centros temáticos.

Para determinar o *status* de cada nodo temático, assume-se que os descritores dos nodos principais devem ser encontrados ao longo de todo o texto. São definidos nodos temáticos 'principais' (com relações textuais com a maioria dos outros nodos principais) e 'específicos' (incluem descritores presentes em, pelo menos, dois diferentes nodos temáticos principais), além de 'conceitos mencionados'.

5.4 *Índice estruturado em árvore binária*

Matsumura, Takasu e Adachi [MAT 00] constroem um índice estruturado, representado por uma árvore binária, para recuperação de textos em língua Japonesa. Essas estruturas são construídas em três etapas: análise morfológica, análise de dependência baseada em padrões e análise de substantivos compostos.

Um relacionamento de dependência é constituído de uma *relation word*, que identifica o relacionamento, e duas *concept words*, que são os argumentos da relação. *Concept words* incluem substantivos, adjetivos, advérbios e constituintes de substantivos compostos. *Relation words* incluem preposições, verbos auxiliares e principais e suas combinações.

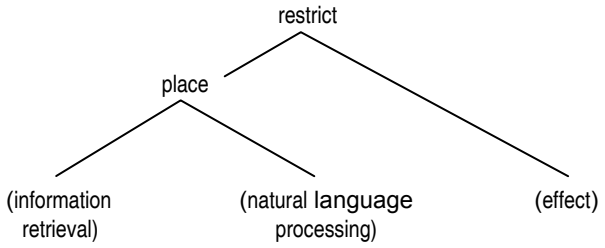

Figura 3. Exemplo de árvore binária

Relation words e *concept words* são extraídas do título e do resumo dos textos, no caso dos documentos, ou diretamente da consulta, expressa na forma de sintagma nominal. A análise de dependência, realizada manualmente, se baseia em padrões. Por exemplo: o padrão 'c1 *of* c2 *on* c3' é substituído por 'c1 *restrict* c2 *place* c3', onde c1, c2 e c3 são *concept words*, e *restrict* e *place* representam *relation words*. A Figura 3 [MAT 00] apresenta, como exemplo, uma árvore binária gerada a partir da sentença em Inglês 'effect of natural language processing on information retrieval'.

A análise dos substantivos compostos possibilita a especificação de relacionamentos entre os conceitos contidos nesses substantivos. É adotado o princípio que estabelece que os substantivos compostos podem ser transformados através da ligação de seus componentes por palavras funcionais. Por exemplo: em Inglês '*information retrieval*' pode ser transformado em '*retrieval of information*' e a análise de dependência pode ser novamente aplicada.

5.5 *Triplas com relações semânticas*

Litkowski [LIT 00] propõe a representação de textos através de triplas com relações semânticas, para melhorar a eficiência de sistemas questão-resposta (do Inglês *question-answering*). A extração das triplas é realizada através dos seguintes passos: identificação das sentenças do texto, análise sintática de cada sentença e análise da árvore sintática para a geração das triplas.

O analisador sintático utilizado adota uma gramática com 350 regras e um dicionário com as categorias gramaticais das palavras.

As triplas, extraídas das árvores sintáticas construídas, são constituídas por uma entidade do discurso, uma relação semântica e uma palavra governante (*governing word*).

Entidades do discurso podem ser de diversas naturezas, como números, seqüências de adjetivos, pronomes possessivos e seqüências de substantivos.

As relações semânticas identificam os papéis semânticos das entidades. Estes papéis são, por exemplo, agente, tema, tempo e modificador. Tais papéis são identificados por termos como SUBJ, OBJ, TIME e ADJMOD. Também são usadas, como identificadores de relações, as preposições que encabeçam os sintagmas preposicionais.

As palavras governantes são, geralmente, aquelas com as quais as entidades do discurso se relacionam na sentença. No caso de SUBJ, OBJ e TIME, as palavras governantes são os verbos principais das sentenças. No caso de preposições, elas são os substantivos ou os verbos que os sintagmas preposicionais modificam. No caso de ADJMOD, elas são geralmente os substantivos modificados.

5.6 *Pares de termos lematizados*

Arampatzis e co-autores [ARA 97, ARA 00, ARA 00a] avaliam diversas abordagens de indexação incluindo termos, bigramas (substantivo-adjetivo) e pares modificado-modificador. O texto é processado através de (i) identificação de sentenças e palavras, (ii)

etiquetagem morfo-sintática, (iii) extração de sintagmas nominais, (iv) eliminação de *stopwords*, (v) transformação de sintagmas nominais de três ou mais palavras para duas palavras, e (vi) normalização morfológica.

A transformação de sintagmas nominais se baseia na freqüência de ocorrência, criando pares de palavras com associação mais freqüente estatisticamente.

Duas abordagens para normalização morfológica foram testadas: lematização e *stemming*.

A diferença entre bigramas e pares modificado-modificador, nesta estratégia, pode ser explicada com o seguinte exemplo. Das frases em Inglês 'air pollution' e 'pollution of the air' seriam extraídos dois bigramas – '(air,pollution)' e '(pollution,air)' –, mas apenas um par modificado-modificador resultante de normalização sintática – '(pollution,air)'.

Levando em conta os resultados obtidos, o processo de lematização se mostrou, segundo Arampatzis e co-autores, mais eficiente que o de *stemming*. Relacionamentos na forma de bigramas tiveram melhores resultados que pares modificado-modificador com consultas curtas; o contrário aconteceu com consultas longas.

5.7 *Expressões ternárias*

Katz e Lin [KAT 00, LIN 01] usam o sistema REXTOR (*Relations EXtracTOR*) para extrair expressões ternárias de textos, aplicáveis em sistema questão-resposta. São utilizadas duas classes de regras para extrair e compor essas expressões: regras de extração e regras de relação. As primeiras são utilizadas para identificar padrões no texto, de acordo com uma gramática. Essas regras descrevem padrões para entidades específicas como sintagmas preposicionais e grupos de substantivos. As regras de relação são ativadas pelas extrações bem sucedidas de entidades específicas. Uma gramática de relações guia a construção de cada expressão ternária.

As expressões ternárias podem ser vistas, de forma intuitiva, como triplas sujeito-relação-objeto. Elas podem ser expressas através de diversos tipos de relações, como relações sujeito-verbo-objeto, relações de posse ou outras. Do ponto de vista sintático, essas expressões podem ser consideradas relações binárias. Do ponto de vista semântico, elas podem ser consideradas predicados com dois argumentos e, assim, podem ser manipuladas, segundo Katz e Lin, através da lógica de predicados.

Na Tabela 2 [KAT 00, LIN 01] são exemplificadas algumas expressões ternárias extraídas de sentenças e frases em Inglês.

Tabela 2. Exemplos de expressões ternárias

texto	expressões ternárias
shiny happy people of Wonderland	<shiny *describes* people> <happy *describes* people> <people *related-to* Wonderland>
the president surprised the country with his actions	<president *is-subject-of* surprise> <country *is-direct-object-of* surprise > <surprise *with* actions>
the meaning of life	<meaning *possessive-relation* life>
the bank near the river	<bank *near-relation* river>

5.8 *Famílias morfológicas e pares de dependência*

Barcala, Vilares e co-autores [BAR 02, VIL 02] usam duas formas para reduzir a variedade lingüística: o uso de morfologia derivacional produtiva e a identificação de pares de termos com dependência sintática. Eles usam o conceito de família morfológica, definida como um conjunto de palavras com mesma raiz. São usados mecanismos de derivação, como derivação de morfemas, identificação de variantes alomórficas (que não apresentam similaridade morfológica), e derivação influenciada por condições fonológicas. Um léxico é utilizado para conferir a existência de cada termo derivado. Cada família morfológica construída possui um termo representante que é utilizado no processo de normalização lexical. Cada termo é normalizado tomando como base a família morfológica à qual pertence. Se uma palavra *p* pertence a uma família morfológica cujo representante é o termo *t*, então *t* será a forma normalizada adotada para *p*.

Os pares de termos com dependência sintática são extraídos com auxílio de gramática. Esses pares podem ser do tipo modificado-modificador, sujeito-verbo ou verbo-complemento, conforme os exemplos em Espanhol '(casa,viejo)', '(perro,comer)' e '(recortar,gasto)', respectivamente.

As expressões em Espanhol 'recorte de gastos' e 'recortar gastos', por exemplo, são normalizadas levando em conta os representantes das famílias morfológicas. Assim '(recorte,gastos)' e '(recortar,gastos)' são normalizadas como '(recorte,gastar)'. Neste exemplo, fica subentendido que 'recorte' é o representante da família de 'recorte' e 'recortar', enquanto que 'gastar' é o representante da família de 'gastos'.

É utilizado o modelo vetorial com esquema TF.IDF tanto para termos quanto para relacionamentos, ou seja, os pares de termos com dependência sintática.

Barcala, Vilares e co-autores relatam haver testado alternativas para normalização lexical, como *stemming*, lematização e o processo proposto, com representante de família morfológica. Os resultados, principalmente quanto à precisão na RI, melhoram quando são usados pares de termos com representantes de famílias morfológicas. Eles concluem que sua proposta favorece a performance de sistemas de RI, principalmente no caso de idiomas morfologicamente ricos como o Espanhol.

5.9 *Bitermos*

Srikanth e Srihari [SRI 02] usam o conceito de bitermo para representar dependência de termos na abordagem probabilística de modelagem de linguagem. De acordo com esses autores, um bitermo é similar a um bigrama exceto pelo fato de a restrição de ordem dos termos ser atenuada, ou seja, bitermos são pares de termos não ordenados. Os relacionamentos são constituídos, então, por bitermos, sendo que um bitermo $\{i,j\}$ corresponde aos bigramas (i,j) e (j,i).

Dos esquemas que os autores desta estratégia testaram, o que produziu melhores resultados de recuperação, considerando o peso de um bitermo $\{i,j\}$, leva em conta a freqüência de ocorrência dos bigramas (i,j) e (j,i) e dos termos componentes, assim como o total de termos do documento corrente e da coleção de documentos.

5.10 *Pares Modificado-Modificador em Conexões Gramaticais*

Changki Lee e Gary Lee [LEE 05] adaptaram uma estratégia desenvolvida por Rijsbergen [RIJ 79] com a inclusão de conexões gramaticais através de árvore de análise sintática com dependências. Rijsbergen adotou o algoritmo proposto por Chow e Liu [CHO 68] para incorporar dependências de termos ao modelo probabilístico. Na abordagem de Rijsbergen é utilizada uma árvore geradora máxima baseada em medida de informação mútua esperada, considerando a distribuição de co-ocorrência dos termos na coleção. Com o uso de árvore de análise sintática, na estratégia de Changki Lee e Gary Lee, é reduzido o número de conexões necessárias e é otimizada a implementação do modelo. Na árvore, o relacionamento entre dois nodos (pai e filho) determina que o nodo filho é dependente (ou modificador) do nodo pai. Na frase, em Inglês, 'brown dog', por exemplo, 'dog' é o nodo pai e 'brown' o nodo filho.

São utilizados somente termos como descritores de conceitos. Os relacionamentos servem apenas para determinar as dependências dos termos. Os pares modificado-modificador são extraídos dos textos e armazenados em uma base de dados. Essa base é pesquisada para verificar se há dependência, ao ser calculado o peso de cada termo, conforme o esquema descrito a seguir.

Para calcular a representatividade de um descritor *i*, é levado em conta o número de documentos onde *i*, *j* e o par modificado-modificador (*i-j*) ocorrem, sendo *j* o modificado e *i* o modificador.

Embora o custo da recuperação aumente devido ao uso da árvore de análise sintática, tanto em fase de indexação (sobre os documentos), quanto em fase de busca (aplicada sobre a consulta), a performance do sistema melhora em relação ao simples uso da fórmula Okapi BM25 (Equação 2).

6 Considerações Finais

A seguir são apresentados na Tabela 3, de forma integrada, dados sobre cada trabalho correlato descrito, quanto aos descritores utilizados e ao esquema adotado para o cálculo da representatividade de cada descritor. Na Tabela 4, o mesmo é feito quanto aos processos para normalização lexical e normalização sintática. Alguns trabalhos descritos apresentam alternativas e, nesses casos, os dados da alternativa de melhor performance são considerados.

Na Tabela 3, a coluna 'descritores' contém informações sobre os tipos de descritores utilizados em cada estratégia. Na coluna 'peso dos descritores' são identificados os principais fatores que influenciam o cálculo da representatividade dos descritores.

Dentre os trabalhos descritos, somente a estratégia com 'Índices Múltiplos' menciona a utilização de arquivos de índice separados para buscas independentes. A estratégia com 'Pares Modificado-Modificador' usa um arquivo de índices, para os termos, e uma base de dados com informações para relacionar termos modificados e modificadores.

Apenas as estratégias com 'Índices Múltiplos' e com 'Árvore Binária' adotam tanto termos quanto relacionamentos como descritores, e especificam cálculo de peso para ambos.

Tabela 3. Análise comparativa dos trabalhos correlatos (I)

estratégia	descritores	peso dos descritores
'Expressões de Índice'	termo-conector-termo	não mencionado
'Índices Múltiplos'	termos, termos compostos, nomes próprios e pares modificado-modificador	esquema TF.IDF com i ncremento de peso para os relacionamentos
'Nodos Temáticos'	descritores principais (ou não) de nodos (temáticos e específicos) além de conceitos mencionados	pesos por classe de descritor
'Árvore Binária'	termos e relacionamentos formados por duas *concept words* e uma *relation word*	peso dos relacionamentos depende do peso dos termos
'Triplas Semânticas'	triplas com relações semânticas	freqüência de ocorrência
'Pares Lematizados'	pares de termos lematizados	esquema TF.IDF
'Expressões Ternárias'	expressões ternárias	freqüência de ocorrência
'Famílias Morfológicas'	pares modificado-modificador	freqüência de ocorrência
'Bitermos'	bitermos	peso do bitermo depende da ocorrência dos termos e dos bigramas
'Pares Modificado-Modificador'	termos	peso dos termos depende das conexões gramaticais

Na Tabela 4, a coluna 'normalização lexical' contém informações sobre o uso de processos de conflação ou procedimentos alternativos para normalização lexical. Na coluna 'normalização sintática' aparecem indicações sobre a abordagem utilizada para a normalização dos relacionamentos.

Somente na estratégia com 'Nodos Temáticos' há normalização léxico-semântica. A estratégia com 'Nodos Temáticos' usa sinonímia através de um thesaurus de domínio. Os outros trabalhos executam somente normalização morfológica e usam lematização ou *stemming*. A estratégia com 'Índices Múltiplos' inclui relações semânticas através de expansão de consulta.

Tabela 4. Análise comparativa dos trabalhos correlatos (II)

estratégia	normalização lexical	normalização sintática
'Expressões de Índice'	sem informação	não
'Índices Múltiplos'	semelhante a *stemming*	regras gramaticais
'Nodos Temáticos'	lematização e sinonímia	não
'Árvore Binária'	lematização*	transformação de substantivos compostos
'Triplas Semânticas'	lematização*	regras gramaticais
'Pares Lematizados'	lematização	apoiada por estatística
'Expressões Ternárias'	lematização*	regras tranformacionais
'Famílias Morfológicas'	com representante de família morfológica	regras gramaticais
'Bitermos'	sem informação	não
'Pares Modificado-Modificador'	lematização*	não

* deduzido dos exemplos apresentados pelos autores.

Dos dez trabalhos descritos, quatro não realizam normalização sintática. Apenas os trabalhos que utilizam regras gramaticais produzem os relacionamentos já normalizados sintaticamente. Os restantes – com 'Pares Lematizados', com 'Expressões Ternárias' e com 'Árvore Binária' – normalizam os relacionamentos após serem identificados.

Referências

[ALL 03] ALLAN, J.; KUMARAN, G. Stemming in the Language Modeling Framework. 26th Annual International ACM SIGIR Conference, 2003. Proceedings, p.455-456.

[ARA 97] ARAMPATZIS, A. T.; KOSTER, C. H. A.; TSORIS, T. IRENA: Information Retrieval Engine based on Natural Language Analysis. Computer-Assisted Information Searching on Internet – RIAO, 1997. Proceedings, p.159-175.

[ARA 00] ARAMPATZIS, A. T.; WEIDE, T. P.; KOSTER, C. H. A.; BOMMEL, P. Linguistically-motivated Information Retrieval. **Encyclopedia of Library and Information Science**, V.69, 2000. p.201-222.

[ARA 00a] ARAMPATZIS, A. T.; WEIDE, T.; KOSTER, C. H. A.; BOMMER, P. An Evaluation of Linguistically-motivated Indexing Schemes. BCS-IRSG – Colloquium on IR Research, 2000. Proceedings, p.91-111.

[BAE 99] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.

[BAR 02] BARCALA, F. M.;VILARES, J.; ALONSO, M. A.. GRAÑA, J.; VILARES, M. Tokenization and Proper Noun Recognition for Information Retrieval. 13th International Workshop on Database and Expert Systems Applications (DEXA), 2002. IEEE Computer Society publication.

[BRA 99] BRASCHLER, M.; RIPPLINGER, B. How Effective is Stemming and Decompounding for German Text Retrieval?. **Information Retrieval Journal**, V. 7, 2004. p.291-316.

[BRU 91] BRUZA, P. D.; WEIDE, Th. P. The Modelling and Retrieval of Documents using Index Expressions. ACM SIGIR Forum, 1991. V.25, N.2, p.91-103.

[CHO 68] CHOW, C.; LIU, C. Approximating discrete probability distributions with dependence trees. **IEEE Translations on Information Theory**, V. IT-14, N. 3, 1968. p.462-467.

[CRO 91] CROFT, W. B.; TURTLE, H. R.; LEWIS, D. D. The use of phrases and structured queries in information retrieval. 14th Annual International ACM SIGIR Conference, 1991. Proceedings, p.32-45.

[DIA 00] DIAS, G.; GUILLORE, S.; BASSANO, J.; LOPES, J. G. Extraction Automatique d'Unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire. *In*: JACQUEMIN, CHRISTIAN (editor). **Traitement Automatique des Langues pour les Recherche d'Information**. Hermès Science Publications, Paris, 2000. p.447-493.

[DOB 98] DOBROV, B.; LOUKACHEVITCH, N. V.; YUDINA, T. N. Conceptual Indexing using Thematic Representation of Text. In: Voorhees, E; Harman, D. K. (editores). The Sixth Text Retrieval Conference (TREC-6), NIST Special Publication, 1998. p. 403-454.

[FAG 87] FAGAN, J. L. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. 10th Annual International ACM SIGIR Conference, 1987 . Proceedings, p.91-101.

[FRA 92] FRAKES, W. B.; BAEZA-YATES. **Information Retrieval: Data Structures and Algorithms**. Prentice-Hall, New York, 1992.

[HIE 00] HIEMSTRA, D. A probabilistic justification for using tfxidf term weighting in information retrieval. **International Journal of Digital Library**, V. 3, Springer-Verlag, 2000. p.131-139.

[JAC 97] JACQUEMIN, C.; KLAVANS, J. L.; TZOUKERMANN, E. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. 35th Annual Meeting of ACL – 8th Conf. of the European Chapter of the ACL, 1997. Proceedings, p.24-31.

[JAC 99] JACQUEMIN, C.; TZOUKERMANN, E. NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. In: Strzalkowski, Tomek (Ed.). **Natural Language Information Retrieval**. Kluwer Academic Publishers, 1999. p.25-74.

[KAT 00] KATZ, Boris; LIN, Jimmy. REXTOR: A System for Generating Relations from Natural Language. ACL'2000 – Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong-Kong, University of Science and Technology, 2000.

[KOR 04] KORENIUS, T.; LAURIKKALA, J.; JÄRVELIN, K.; JUHOLA, M. Stemming and Lemmatization in the Clustering of Finnish Text Documents. 13th ACM Conference on Information and Knowledge Management (CIKM), 2004. Proceedings, p.625-634.

[KRO 93] KROVETZ, R. Viewing Morphology as an Inference Process. 16th Annual International ACM SIGIR Conference, 1993. Proceedings, p.191-202.

[LEE 05] LEE, C.; LEE, G. G. Probabilistic information retrieval model for a dependency structured indexing system. Information. **Processing and Management**, V. 41, 2005. p.161-175.

[LIN 01] LIN, Jimmy. Indexing and Retrieving Natural Language using Ternary Expressions. Dissertação de Mestrado, Massachusetts Institute of Technology, Cambridge, EUA, 2001.

[LIT 00] LITKOWSKI, K. Question-Answering Using Semantic Relation Triples. In: VOORHEES, E; HARMAN, D. (editores). The Eighth Text Retrieval Conference (TREC-8), NIST Special Publication, Gaithersuburg, 2000. p.349-356.

[LOU 99] LOUKACHEVITCH, N. V.; SALLI, A. D.; DOBROV, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. NLDB'99 – 4th Int. Conf. on Applications of Natural Language to Information Systems, 1999. OCG Schriftenreihe, Lecture Notes, V.129, p.203-208.

[LOU 00] LOUKACHEVITCH, N. V.; DOBROV, B. V. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. **Machine Translation Review**, ISSN 1358-8346, N. 11, 2000. p.10-20.

[MAT 00] MATSUMURA, A.; TAKASU, A.; ADACHI, J. The Effect of Information Retrieval Method Using Dependency Relationship Between Words. RIAO'2000 – Multimedia Information Representation and Retrieval, 2000.

[MAR 03] MARTINS, R.; NUNES, Graça; HASEGAWA, R. Curupira: A Functional Parser for Brazilian Portuguese. PROPOR, 2003. p.179-183.

[RIJ 79] RIJSBERGEN, C.J. **Information Retrieval**. London: Bitterworths, 1979.

[ROB 94] ROBERTSON, S. E.; WALKER, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: Croft, W. B.; van Rijsbergen, C. J. (Eds.). 17th Annual International ACM SIGIR Conference, 1994. Proceedings, p.232-241.

[SAL 83] SALTON, G.; MACGILL, M. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

[SAL 88] SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing and Management**, V. 24, N. 5, 1988. p.513-523.

[SAV 03] SAVARY, A., and C. JACQUEMIN. Reducing Information Variation in Text. Text- and Speech-triggered Information Access, Springer, Lectures Notes in Artificial Intelligence 2705, 2003. p.145-181.

[SPA 00] SPARCK-JONES, K. ; WALKER, S. ; ROBERTSON, S. E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments – Part 1 and 2. Information Processing and Management, V. 36, N. 6, 1997. p. 779-840.

[SRI 02] SRIKANTH, M.; SRIHARI, R. 2002. Biterm language models for document retrieval. 25th Annual International ACM SIGIR Conference, 2002. Proceedings , p.425-426.

[STR 96] STRZALKOWSKI, T.; PEREZ-CARBALLO, J.; MARINESCU, M. Natural Language Information Retrieval in Digital Libraries. First ACM International Conference on Digital Library, 1996. p.117-125.

[STR 99] STRZALKOWSKI, T.; LIN, F.; WANG, J.; PEREZ-CARBALLO, J. Evaluating Natural Language Processing Techniques in Information Retrieval. *In*: Strazalkowski, T. (ed.) **Natural Language Information Retrieval**. Kluwer Academic Publishers, 1999. p.113-145.

[VIE 02] VIEIRA, R.; SALMON-ALT, S.; SCHANG, E. Multilingual Corpora Annotation for Processing Definite Descriptions. Advances in NLP, 3th International Conference, PorTal, 2002. Proceedings, p.249-258.

[VIL 02] VILARES, J., BARCALA, F. M.; ALONSO, M. A. Using Syntactic dependency-pairs conflation to improve retrieval performance in Spanish. **Computational Linguistics and Intelligent Text Processing**, Springer-Verlag, Lectures Notes in Computer Science, 2276, 2002. p.381-390.

[WON 00] WONDERGEM, B.; BOMMEL, P.; WEIDE, Th. P. Matching Index Expressions for Information Retrieval. **Information Retrieval**, V. 2, 2000. p.337-360.

[WON 00a] WONDERGEM, B.; BOMMEL, P.; WEIDE, Th. P. Nesting and Defoliation of Index Expressions for Information Retrieval. **Knowledge and Information Systems**, V. 2, N. 1, 2000.