

Análise e representação de assuntos em sistemas de recuperação da informação; linguagens de indexação

Subject analysis and indexing languages in information retrieval systems

MARIA CRISTINA MELLO FERREIRA PINTO *

A subjetividade interfere no processo de indexação em dois pontos: na determinação do conteúdo informativo dos documentos e na escolha de termos indexadores representativos deste conteúdo. O conhecimento da estrutura das linguagens de indexação e seus componentes básicos (vocabulário e sintaxe) ajuda a aumentar a objetividade neste processo reduzindo a subjetividade a um mínimo aceitável. O controle de vocabulário exercido em sistemas de recuperação da informação mostra as relações existentes entre os termos indexadores ajudando os usuários a montarem estratégias de busca mais eficientes. Para o bibliotecário o estabelecimento destas relações facilita a escolha dos termos indexadores mais adequados à representação de assunto dos documentos.

* Professora da Escola de Biblioteconomia da UFMG.

INTRODUÇÃO

Muito se tem escrito sobre linguagens de indexação, particularmente sobre a análise de assuntos de documentos, em termos da representação destes assuntos de acordo com algum padrão. No entanto, pouco se tem visto sobre o processo de análise de assunto em si, como um processo mental de determinação do conteúdo informativo de documentos. Este artigo pretende discutir justamente esse aspecto, como também apresentar uma nova abordagem para o estudo das relações entre os termos e conceitos em linguagens de indexação, visando facilitar uma metodologia mais racional no processo de análise de assunto com vistas à recuperação da informação. Conceitua-se linguagem de indexação e estudam-se seus elementos componentes: o vocabulário e a sintaxe. O vocabulário, numa linguagem de indexação, evidencia as relações existentes entre os termos indexadores. O estudo da sintaxe mostra a possibilidade da coordenação de termos indexadores para a representação de assuntos compostos nesse tipo de linguagem. Considera-se a influência das estruturas de sistemas pré e pós-coordenados em relação ao problema sintático de citação de assuntos compostos.

1. INDICES E INDEXAÇÃO

Um índice é o instrumento mais importante num sistema de recuperação de informações. Tem sido descrito como a chave para o acesso ao conteúdo dos documentos, ou como uma ponte ou elo entre o conteúdo de um arquivo de informações e seus usuários. A indexação consiste na indicação do conteúdo informativo de um documento através da determinação de um ou mais termos que representarão esse conteúdo. Essencialmente a indexação inclui dois estágios:

- a) Estabelecimento dos conceitos tratados e seleção dos que serão indexados, tendo em vista os objetivos do sistema;
- b) Tradução dos conceitos selecionados em termos indexadores de acordo com algum padrão consistente, ou seja, de acordo com a linguagem de indexação usada no sistema.

Podemos dizer que as linguagens de indexação atuam nos sistemas de recuperação de informações em dois níveis: orientando o indexador sobre quais os melhores termos para a representação do assunto de um documento e orientando o pesquisador sobre a escolha dos termos indexadores que corresponderiam à representação do assunto por ele procurado. Isto é, orientaria a elaboração de uma estratégia de busca que o levasse com menor esforço e maior eficiência aos documentos procurados. Para que esse segundo nível seja atingido é necessário que no momento da indexação (primeiro nível) a linguagem não somente oriente o indexador sobre os termos indexadores a serem usados, como também o instrua sobre quais orientações deve estabelecer no catálogo (índice) para guiar o usuário na elaboração de sua estratégia de busca. Essas orientações vão desde o simples controle de termos sinônimos, até a construção de complexas redes de referências cruzadas, dando ao usuário outras opções de busca em relação ao assunto procurado.

2. ESTRUTURA DAS LINGUAGENS DE INDEXAÇÃO

Qualquer conjunto de termos e/ou símbolos usados para denominar classes de assuntos em sistemas de recuperação da informação constitui uma linguagem de indexação. Como qualquer língua, uma linguagem de

→
vocabulário
→
sintaxe
→
Se se adotam como termos indexadores os termos da forma como aparecem nos documentos sem modificações, está se usando a linguagem natural. No entanto, na recuperação de informações, esse procedimento pode causar vários problemas, como no caso de diferentes autores usando palavras diversas para expressarem o mesmo conceito (dispersão terminológica) ou o caso do uso de diferentes estruturas para expressarem a mesma idéia (dispersão sintática). Por essas razões, a maioria dos sistemas tem adotado algum tipo de controle sobre os termos usados. Esse controle deve agir sobre o vocabulário (termos propriamente ditos) e sobre a sintaxe (estrutura frásica). O controle sobre o vocabulário evidenciará as relações semânticas, ou de significado, que existem entre os termos indexadores. O controle sobre a sintaxe mostrará as relações estruturais ou frásicas entre os conceitos, evidenciando a possibilidade da combinação de elementos para expressar assuntos compostos.

indexação é formada por dois componentes básicos: o **vocabulário** e a **sintaxe**.

Da mesma forma que numa língua o vocabulário é a relação das palavras utilizadas, numa linguagem de indexação o vocabulário corresponde à relação de termos e/ou símbolos empregados para a identificação temática de documentos. Sendo a sintaxe a estrutura gramatical de uma língua, abrange o conjunto de regras relativas à combinação de elementos, que juntos possuem significado diferente daquele que apresentavam originalmente. Numa linguagem de indexação, a sintaxe refere-se ao conjunto de regras para a combinação de termos e/ou símbolos, a fim de expressarem, com a menor ambigüidade possível, o conteúdo temático de documentos. Pode estabelecer regras sobre o uso dos termos indexadores, além de revelar as relações existentes entre os conceitos e os termos da linguagem de indexação.

Se se adotam como termos indexadores os termos da forma como aparecem nos documentos sem modificações, está se usando a linguagem natural. No entanto, na recuperação de informações, esse procedimento pode causar vários problemas, como no caso de diferentes autores usando palavras diversas para expressarem o mesmo conceito (dispersão terminológica) ou o caso do uso de diferentes estruturas para expressarem a mesma idéia (dispersão sintática). Por essas razões, a maioria dos sistemas tem adotado algum tipo de controle sobre os termos usados. Esse controle deve agir sobre o vocabulário (termos propriamente ditos) e sobre a sintaxe (estrutura frásica). O controle sobre o vocabulário evidenciará as relações semânticas, ou de significado, que existem entre os termos indexadores. O controle sobre a sintaxe mostrará as relações estruturais ou frásicas entre os conceitos, evidenciando a possibilidade da combinação de elementos para expressar assuntos compostos.

3. RELAÇÕES SEMÂNTICAS

Essas relações devem ser controladas ou mostradas nas linguagens de indexação com o objetivo de se indicarem os termos alternativos ou substitutivos de indexação e busca. Evidenciam-se na própria definição do termo e são permanentes. Podemos, por exemplo, definir «alumínio» como um «metal não-ferroso» evidenciando-se uma relação permanente entre os dois conceitos: o alumínio será sempre um metal não ferroso. Podem-se subdividir as relações semânticas em 3 subtipos: de equivalência, hierárquicas e afinitivas ou associativas. Quando a linguagem mostra as relações de equivalência fornece termos substitutivos de indexação e busca, que podem ser de várias espécies:

- sinônimos e antônimos (adubos x fertilizantes)
(diminutivos x aumentativos)
- quase sinônimos (datiloscopia x impressão digital)
- grafias diferentes (contato x contacto)
- abreviaturas e acrônimos (ONU x Organização das Nações Unidas)
- traduções (recall x revocação)
- ordem de citação (Automação de bibliotecas x Bibliotecas, automação de).

Por outro lado, a indicação das relações hierárquicas e afinitivas permite a busca por termos alternativos. As relações hierárquicas exprimem o posicionamento de um conceito em relação a termos mais amplos ou mais específicos. As relações afinitivas revelariam termos indexadores que desencadeariam uma ampliação da estratégia de busca. Mostrariam termos afins, onde «um lembra o outro» sem, no entanto, expressarem uma noção de hierarquia. Essas relações são as mais difíceis de serem

estabelecidas e mostradas quando da construção de uma linguagem de indexação.

Podem-se estabelecer relações hierárquicas de dois tipos:

- gênero x espécie

}	quase-genérico (gato x animal de estimação)
}	verdadeiramente genérica (gato x mamífero)
- todo x parte (aparelho respiratório x pulmão)

Para se determinar o tipo de relação hierárquica existente entre termos poder-se-ia acrescentar ao termo mais específico as frases: «uma espécie de» ou «uma parte de» que evidenciaríamos no caso uma relação genérica ou partitiva respectivamente:

o gato é «uma espécie de» mamífero?
o pulmão é «uma parte do» aparelho respiratório?

As relações afinitivas possibilitariam a busca de termos alternativos em relação a:

- coordenação (portas x janelas) (Terra x Marte)
- genética (pai x filho)
- concorrente (educação x ensino)
- causa e efeito (ensino x aprendizagem)
- instrumental (medicina preventiva x vacinação)
- material (esquadrias x alumínio)
- similaridade de processo (classificação x indexação).

Ao se julgar o estabelecimento das relações hierárquicas e afinitivas deve-se ter em mente a necessidade de mostrar somente aquelas que teriam uma utilidade

em potencial, tendo em vista a recuperação da informação. Em termos práticos, essa seleção poderia ser orientada pela questão: uma pessoa interessada no conceito X poderia se interessar pelo conceito Y? Ou seja, uma pessoa interessada em «medicina preventiva» poderia se interessar por «vacinação?» Quando a linguagem mostra as relações hierárquicas e afinitivas possibilita ampliar a pesquisa para termos mais amplos, mais específicos e associados através de uma referência possivelmente do tipo «ver também». Alguém interessado em «metal não-ferroso» poderia pesquisar por «alumínio». Alguém interessado em «classificação» poderia também se interessar por «indexação».

Dependendo da estrutura da linguagem de indexação (alfabética ou classificada) o estabelecimento das relações semânticas pode ser mais complexo ou mais simples. Num arranjo alfabético é imprescindível, em virtude da dispersão de assuntos relacionados pela ordenação alfabética. As relações de equivalência, também chamadas one-way seriam mostradas através de:

[see	
	ver	Exemplo: ONU ver Organização das Nações
	use	urânio...

As relações hierárquicas e afinitivas, também chamadas two-way seriam mostradas através de

see also]	em listas de cabeçalhos de assunto
ver também		
XX		

BT, NT, RT { em tesouros

Exemplo: metal não-ferroso, **ver também:**

alumínio

urânio.

Um sistema desse tipo exige uma rede de referências cruzadas bastante complexa. Num sistema de arranjo sistemático a estrutura de referências é bastante diminuída, pois a própria notação simbólica muitas vezes expressa as relações pela justaposição de elementos. As relações hierárquicas seriam mostradas pelo acréscimo de elementos, como por exemplo na CDD (Classificação Decimal de Dewey) onde:

553.4 = Metais não-ferrosos

553.492 = alumínio



As relações de coordenação mostrariam elementos de uma mesma classe:

553.492 = alumínio

553.493 = urânio

As relações de equivalência seriam mostradas pelo próprio índice alfabético do sistema, ou em notas nas tabelas tais como: «classifique aqui X, Y, Z».

4. RELAÇÕES SINTATICAS

Examinando documentos com o objetivo de se determinar seu conteúdo, pode-se observar que muitas vezes esse conteúdo refere-se a conceitos de natureza e complexidade diversas. Por exemplo, documentos sobre:

conceito	natureza
trabalhador metalúrgico	personalidade
usina siderúrgica	entidade
alto forno	parte
alumínio	material
fundição	ação, operação
ferrugem	ação, processo
flexibilidade	propriedade, qualidade
normas	forma
Minas Gerais	local
1970	tempo, época

Pode ocorrer ainda que o conteúdo não se refira a um conceito único, mas à combinação de conceitos simples ou primários que corresponderiam à representação de um assunto composto:

esquadrias de alumínio	material + material
fundição do alumínio	operação + material
normas para fundição do alumínio	forma + operação + material
ferrugem em esquadrias metálicas	processo + material
usinas siderurgicas de M. Gerais	entidade + local

As relações sintáticas entre os termos de uma linguagem de indexação originam-se justamente da necessidade de se poder recuperar a interseção entre duas ou mais classes de conceitos distintos. Pode existir um interesse por «fundição do alumínio» e não por «fundição» e «alumínio». Essas relações constituem-se da associação de termos para representar assuntos compostos que geralmente aparecem quando se analisam os documentos: são dependentes dos documentos e por isso são tempo-

rárias, ao contrário das relações semânticas que são permanentes. Considerando-se o estado atual do desenvolvimento da ciência e da tecnologia, pode-se notar uma tendência à especialização e, conseqüentemente, a literatura sobre os diversos assuntos espelha essa especialização. Atualmente recebem-se com mais freqüência documentos a serem indexados sobre assuntos compostos, do que sobre assuntos simples. É mais provável receber-se um documento sobre «normas para a fundição de esquadrias de alumínio» do que sobre «alumínio» sob um aspecto geral. A complexidade dos assuntos tem sido grandemente discutida em artigos sobre análise de assunto e linguagens de indexação. Discute-se, por exemplo, até que ponto os conceitos simples são realmente únicos. O conceito «esquadria metálica», por exemplo, se refere a um conceito único, ou envolve dois elementos: «esquadrias» e «metal»? As orientações a respeito desse tipo de julgamento são pouco elucidativas. A complexidade dos assuntos está intrinsecamente relacionada com os aspectos de pré e pós-coordenação de conceitos na recuperação de informações.

4.1 Pré e pós-coordenação

4.1.1 Pré-coordenação

Os chamados sistemas tradicionais de recuperação da informação, como a grande parte das bibliotecas que não usam nenhum grau de automação, criam catálogos de assunto (alfabéticos ou classificados) onde a representação do conteúdo do documento revela a totalidade do seu assunto. Através de um enunciado, muitas vezes único, os vários tópicos relativos ao conteúdo conceitual do documento são reunidos numa entrada que resume todos os aspectos que compõem o assunto. Nesse caso depara-se com o sério problema de se julgar a impor-

tância de cada um desses tópicos, para criar uma ordem de citação que seria ideal para os usuários. Esses sistemas, chamados pré-coordenados, estabelecem a coordenação dos vários tópicos referentes a um assunto composto no momento de indexação, e com grande subjetividade determinam uma prioridade na citação desses elementos. Obviamente os elementos que não o primeiro estariam ocultos, ou seja, o usuário não teria acesso a eles diretamente. Esse tipo de sistema requer o que se chama de «entrada múltipla» ou uma rede de referências ou remissivas bem elaborada. Pode-se usar a rotação dos elementos de modo que cada um passe a ser um ponto de acesso:

Normas para fundição de esquadrias de alumínio
Fundição de esquadrias de alumínio, Normas para
Esquadrias de Alumínio, Normas para fundição de
Alumínio, Normas para fundição de esquadrias de

Esse procedimento ampliaria a possibilidade de acesso mas sua adoção acarretaria várias dificuldades: o sistema ficaria mais caro, o tempo de indexação seria maior, aumentar-se-ia o tamanho do índice, etc.

Desde os tempos de Cutter, quando em 1876 publica sua obra «Rules for a dictionary catalog», existe a preocupação de se orientar o indexador sobre o julgamento da importância dos itens num enunciado de assunto composto. Sua recomendação «somente quando o segundo termo de um cabeçalho composto for notoriamente o mais significativo, a ordem da linguagem natural seria abandonada»... mostra sua tentativa de sistematizar a entrada de assuntos compostos. No entanto, a expressão «o mais significativo» revela um elevado grau de subjetividade. Kaiser, em 1911, na sua obra «Systematic indexing» tenta dar uma certa consistência à orientação

da citação de assuntos compostos: «o concreto, depois o processo». Foi Coates, em 1960, que, aprofundando os estudos sobre o assunto, questionou sobre o motivo pelo qual Kaiser teria usado o princípio do «concreto, depois o processo». Com isso ele chega à sua teoria da concretividade decrescente, propondo:

COISA depois a AÇÃO

Introduzindo o elemento «material»:

COISA — MATERIAL — AÇÃO e

dependentes da coisa: parte, tipo ou propriedade

COISA — PARTE — MATERIAL — AÇÃO

A análise de assunto pelos princípios de Coates envolveria a determinação das partes de um enunciado composto e sua caracterização. Considerando o assunto:

«Fabricação de sacos multifolhados de papel Kraft para a embalagem de cimento», pode-se dividi-lo em duas partes:

a) Fabricação de sacos multifolhados de papel Kraft

↓ ↓ ↓
ação coisa propriedade

b) Embalagem de cimento

↓ ↓
ação coisa

Segundo as orientações de Coates, a ordem de citação seria:

- a) Sacos - papel - Kraft - multifolhados - fabricação
- b) Cimento - embalagem

Coates não deixa claro suas considerações sobre a determinação de qual das partes do enunciado (a) ou (b) deveria ser citada em primeiro lugar. Para se julgar a prioridade dessa citação poder-se-ia perguntar «do que trata esse documento»?

— Embalagem de cimento/através/de sacos multifolhados... A primeira parte desse enunciado seria o assunto e a segunda um aspecto, nesse caso, um instrumento. Assim a citação, segundo Coates, seria:

Cimento - embalagem - sacos - papel - kraft - multifolhado - fabricação.

Esse procedimento não exclui a subjetividade na determinação da importância dos itens (a) e (b). Poder-se-ia considerar que, em determinadas circunstâncias, para usuários interessados em fabricação de sacos multifolhados de papel Kraft a parte **b** deveria ser citada em primeiro lugar.

Obviamente, como já se viu, os elementos que não o primeiro (embalagem, sacos, papel, kraft, multifolhado, fabricação) estariam ocultos num sistema pré-coordenado. Uma entrada única para esse documento limitaria sua recuperação pelo termo «cimento». Coates reforça a idéia de Ranganathan propondo a utilização do procedimento em cadeia para o acesso aos termos ocultos pela estrutura sintática:

Fabricação - multifolhado - Kraft - papel - sacos - embalagem - cimento **ver**
cimento - embalagem - sacos - papel - Kraft
- multifolhado - fabricação.

Multifolhado - Kraft - papel - sacos - embalagem - cimento, **ver**
cimento - embalagem sacos - papel - kraft
- multifolhado.

Kraft - papel - sacos - embalagem - cimento, **ver**
cimento - embalagem - sacos - papel - kraft.

Papel - sacos - embalagem - cimento, **ver**
cimento - embalagem - sacos - papel.

Sacos - embalagem - cimento, **ver**
cimento - sacos - embalagem.

Embalagem - cimento, **ver**
cimento - embalagem.

Esse procedimento, como já vimos, onera o sistema dificultando sua operação. No entanto os princípios de Coates foram e são válidos para o estudo das relações sintáticas em linguagens de indexação, constituindo uma sistemática para a análise de assuntos compostos. Considerando a subjetividade inerente a todo o processo de indexação e análise de assunto, os princípios de Coates permanecem como uma orientação para se tentar reduzir essa subjetividade ao mínimo possível.

4.2 Pós-coordenação

Os sistemas mais modernos de recuperação da informação, notadamente os sistemas automatizados, transferiram a coordenação dos elementos básicos de um

assunto composto do momento da indexação para o momento de busca. Por isso são chamados pós-coordenados: pós-coordenação em relação ao momento da indexação. Nesse tipo de sistema não existe a preocupação com a importância dos elementos de um assunto composto e como consequência com a ordem de citação desses elementos. São também chamados de «sistemas invertidos» (**inverted file**) porque neles o catálogo de assunto, onde efetivamente se opera a busca de documentos, é formado por entradas que representam os assuntos contidos na coleção. Nos sistemas tradicionais esses catálogos são formados por entradas que representam os documentos. Inversamente, nos sistemas pós-coordenados o catálogo de busca registra os assuntos, relacionando a eles um código qualquer que represente cada documento que contenha aquele assunto. No momento da busca examina-se o arquivo, independentemente da ordem de citação. Um sistema pós-coordenado consiste de entradas que normalmente são conceitos únicos (termos indexadores) e de uma saída que permite a comparação das entradas de assunto, com o objetivo de determinar coincidências que revelem documentos pertinentes ao assunto composto pesquisado.

Tomando o exemplo de Coates «Fabricação de sacos multifolhados de papel Kraft para embalagem de cimento», num sistema pós-coordenado dá-se entrada a todos os termos relevantes, sem a necessidade de decisões a respeito da importância dos itens: fabricação, sacos, multifolhados, papel, kraft, embalagem, cimento. Num sistema pré-coordenado criou-se uma entrada única para este todo, através do estabelecimento da importância dos itens. Num sistema pós-coordenado a indexação desse documento gerou a criação de sete entradas. Isto não quer dizer que, usando um sistema pós-coordenado, não haja qualquer problema em relação ao controle de

vocabulário. Os problemas relativos ao controle semântico e sintático do vocabulário permanecem.

O problema semântico relativo às relações de equivalência, hierárquicas e afinitivas, deverá ter uma solução dentro das mesmas diretrizes usadas num sistema pré-coordenado: o uso de remissivas para relações de equivalência e o uso de referências do tipo «ver também» para relações hierárquicas e afinitivas. O problema sintático, relativo aos assuntos compostos, apesar de simplificado, em relação à ordem de citação, permanece no que concerne à determinação das partes componentes de um assunto composto. Discute-se até que ponto um enunciado do tipo «trabalhador metalúrgico de alto forno» envolve realmente duas partes «trabalhador metalúrgico» e «alto forno» ou corresponderia a um conceito único, se referindo a um tipo de trabalhador com determinadas qualificações. A fatoração de um termo composto em suas partes componentes pode, às vezes comprometer o sentido do termo. Os termos «canção-de-roda», «copo-de-leite», «carro-biblioteca» se fatorados exprimiriam conceitos diferentes do que teriam se pré-coordenados. Outro problema que pode ocorrer seria o relativo à mudança do sentido dos termos ao serem coordenados no momento da busca, pelas funções que exercem na frase. Como por exemplo: «filosofia da história» ou «história da filosofia», «dosagem de hidrocarbonatos na água» ou «dosagem da água nos hidrocarbonatos», etc.

5. CONCLUSÕES

Na prática, nota-se uma tendência ao uso de algum tipo de controle de vocabulário nos sistemas de recuperação da informação e à criação de instrumentos de indexação adequados às peculiaridades dos sistemas. As listas de cabeçalhos de assunto têm sido criadas e usadas

em sistemas pré-coordenados e os tesauros constituem a linguagem adotada em sistemas pós-coordenados. Os tesauros exercem um controle menos rígido, mais sugestivo e flexível. Utilizam um grau de pré-coordenação mínimo, somente aquele necessário ao não comprometimento da significação dos descritores.

Analisar assuntos contidos em documentos, visando a sua recuperação quando necessário é uma das tarefas mais difíceis para o bibliotecário. Talvez isso se explique pela subjetividade inerente ao processo. É desejável o conhecimento do potencial interesse dos usuários do sistema, podendo-se dizer que o conhecimento da técnica da análise de assunto é imprescindível. O bibliotecário deverá estar apto a saber caracterizar as partes de um enunciado composto, escolher uma ordem lógica para sua citação e criar, no catálogo, orientações para os usuários do sistema. Reconhece-se a vastidão e a complexidade do procedimento da análise e representação de assuntos em sistemas de recuperação de informação e espera-se que este artigo tenha contribuído de alguma forma para elucidar os problemas relativos a esta área tão discutida e de soluções tão questionáveis.

The indexing process implies a great degree of subjective judgment, mainly in connection with identifying information content of documents and with choosing indexing terms which represent this content. The knowledge of indexing languages structure and their basic components (vocabulary and syntax) helps to increase objectivity in that process. Vocabulary control in information retrieval systems shows the existing relationship among indexing terms, thus helping users to develop more efficient search strategies. It also helps librarians to choose more appropriate indexing terms to represent the subject content of documents.

REFERÊNCIAS BIBLIOGRÁFICAS

1. CLEVERDON, Cyril. Evaluation tests of information retrieval systems. *Journal of Documentation*, London, **26**(1): 55-57, Mar. 1970.
2. CLEVERDON, Cyril. The Cranfield hypothesis. *The Library Quarterly*, Chicago, **35**(2): 121-4, Apr. 1965.
3. COATES, E. J. Classification in information retrieval: the twenty years following Dorking. *Journal of Documentation*, London, **34**(4): 288-99, Dec. 1973.
4. FOSKETT, A. C. *The subject approach to information*. 4. ed. London, Clive Bingley, 1982. 574 p.
5. LANCASTER, F. Wilfrid. *Vocabulary control for information retrieval*. Washington, Information Resources, 1972. 233 p.
6. NEEDHAM, C. D. *Organizing knowledge in libraries; an introduction to information retrieval*. 2. ed. London, A. Deutsch, 1971. 448 p.
7. VICKERY, B. C. *Classificação e indexação nas ciências*. Rio de Janeiro, BNG/Brasilart, 1980. 274 p.