Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação

Jaime Robredo
Murilo Bastos da Cunha

Resumo

A análise das coocorrências entre pares de palavras permite estabelecer índices estatísticos que representam a força de associação entre esses pares e, a partir dos valores encontrados, mapear o estado de uma área do conhecimento num determinado momento. A identificação de aglomerados de palavras-chave e a análise da força de ligação entre pares de palavras e expressões significativas integrantes dos aglomerados abre o caminho para importantes aplicações que vão da construção de léxicos especializados até o desenvolvimento de instrumentos lógicos suscetíveis de otimizar os processos de indexação automática e recuperação da informação, passando pela possibilidade de acompanhar a evolução dos temas de interesse da pesquisa científica. Apresenta-se uma aplicação da análise das coocorrências de pares de palavras-chave para identificação do âmbito e da abrangência do léxico básico, que caracteriza os processos de indexação e recuperação da informação.

Palavras-chave

Léxico básico; Indexação; Recuperação da informação; Técnicas infométricas.

INTRODUÇÃO

Em um artigo que pode ser considerado como um clássico na matéria, Whittaker[1] definia, em 1989, a análise das coocorrências das palavras (em inglês, *co-word analysis*) "como a utilização do 'comportamento' das palavras como um meio para elucidar as estruturas das idéias e outros problemas representados em conjuntos adequados de documentos". Essa definição encontra seu fundamento, de acordo com o mesmo autor, nos seguintes princípios:

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

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

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

Se os pressupostos acima fazem sentido, nada impede utilizar as freqüências com que ocorrem os possíveis pares de palavras relacionadas, em cada artigo integrante de um conjunto de artigos, como um meio para descrever a estrutura dos conceitos contidos nos artigos.

O mesmo autor acrescenta uma quarta premissa: "que as palavras-chave escolhidas por indexadores competentes como descritores do conteúdo dos artigos são de fato uma indicação confiável dos conceitos científicos a que se referem, o que torna possível o uso das palavras-chave como o elemento base para a análise das coocorrências das palavras".

Mediante a análise das coocorrências entre pares de palavras, é possível estabelecer índices estatísticos que representam a 'força' de associação entre esses pares e, a partir dos valores encontrados, elaborar diversos tipos de representações gráficas (árvores, redes, agrupamentos diversos) e, assim, visualizar (ou, utilizando um anglicismo bem em voga, 'mapear') o estado de um campo do conhecimento, em um determinado momento.

Uma descrição bastante completa do desenvolvimento dos métodos de análise das coocorrências entre pares de palavras, até 1986, pode-se encontrar na obra de Callon, Law e Rip[2]. Uma importante bibliografia mais atualizada encontra-se em uma recente comunicação de um dos autores do presente trabalho[3]. Dentre os numerosos autores que nos últimos anos aplicaram a análise das coocorrências de palavras-chave ao estudo da situação ou da evolução de diversas áreas da ciência, merecem destaque, além de Whittaker[1], já citado, King[4], Law e Whittaker[5], Leydesdorff[6], Callon, Courtial e Laville[7].

Neste trabalho, apresenta-se uma aplicação da análise da coocorrência de pares de palavras-chave para identificação do âmbito e da abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação.

O objetivo da pesquisa é mostrar a possibilidade de identificar agrupamentos de palavras-chave que caracterizam determinados conceitos básicos de um campo específico. Em outras palavras, trata-se de uma tentativa de descer a um nível de detalhamento maior (identificação de agrupamentos e/ou cadeias de termos significativos inter-relacionados suscetíveis de caracterizar, de *per se* ou em conjunto, uma determinada área de conhecimento) do contemplado por Diodato [8], quando visualiza simplesmente a análise de coocorrências como um método de identificar, num determinado acervo, grupos de documentos que possuem certa afinidade no seu conteúdo.

METODOLOGIA

O *corpus* de termos e expressões significativas utilizado no presente estudo foi extraído do manuscrito da obra *Glossário de Termos Técnicos de Ciência da Informação* [9], mediante um processo de indexação automática de 222 verbetes relacionados com os diversos aspectos da indexação e recuperação da informação, utilizando o sistema InfoDoc[10],[11]. A indexação automática rendeu 26 278 entradas no índice, com freqüências variando entre 1 (15 979 termos) e 95 (1 termo).

Para o estudo dos agrupamentos binários, foi aplicada a equação se-guinte, utilizada por diversos autores: [12-18]:

$$Eij = (Fij)^2 / Fi \, . \, Fj$$

onde: Eij (coeficiente de equivalência) é um índice que mede a 'força' ou probabilidade de associação (coocorrência) dos termos *i* e *j* no conjunto de verbetes, *Fi* e *Fj* são, respectivamente, as freqüências (ocorrências) dos termos *i* e *j*, e *Fij* é a freqüência com que o par de termos *i* e *j* aparecem juntos (coocorrência) nos diversos verbetes.

O cálculo de *Eij* foi realizado utilizando um programa, especialmente desenvolvido para este estudo, o qual recebe como entrada a relação de termos considerados significativos pelo InfoDoc®, no processo de indexação automática, com suas respectivas freqüências de aparecimento no conjunto de verbetes (tabela 1, a seguir) e gera uma tabela que indica, para os pares de termos (*Fi* e *Fj*) que ocorrem com freqüência igual ou maior que 2, os valores correspondentes à freqüência de associação ou total de co-ocorrências (*Fij*), no conjunto de verbetes, e ao coeficiente de equivalência *Eij* (tabela 2, a seguir).

Observe-se que o valor de *Eij* é 1 (um) quando a ocorrência de *i* implica a ocorrência de *j* e vice-versa. Inversamente, seu valor é 0 (zero) quando a presença de um dos termos exclui a ocorrência do outro, ou seja, nenhum verbete é indexado simultaneamente pelos dois termos.

A observação dos fragmentos da tabela de ocorrências e coocorrências, representada na tabela 2, permite ver que, quando o número de coocorrências é menor que 2 ($Fij < 2$) e o número de ocorrências dos termos que integram o par considerado é muito díspar (por exemplo: $Fi = 11$ e $Fj = 59$; $Fi = 11$ e $Fj = 54$, etc.), o valor de *Eij* é inferior a 0,01 ($Eij < 0{,}01$) e aparece na tabela como 0.00, indicando que a possibilidade de coexistência dos dois termos do par é praticamente nula.

Por essas razões, aplicando um critério semelhante ao já aplicado anteriormente por Polanco [14-16] e, mais recentemente por Basevi[17] e Lima[18], que consiste em eliminar os termos e expressões de baixa freqüência (muito numerosos) e os de freqüência muito elevada (pouco numerosos), foram eliminados para formar os aglomerados ou agrupamentos (em inglês, *clusters*) os termos com freqüência inferior a 2 ou muito elevada (termos muito genéricos, ou obviamente redundantes com o tema escolhido para estudo; por exemplo, *assunto, documento, informação, item, palavra, termo* etc.)*.

Desta forma, foram retidos em primeira instância 381 termos que, com os termos a eles associados constituiriam possíveis agrupamentos. Após reunir em uma só entrada os termos ou expressões que são simples variações mórficas de um mesmo conceito (por exemplo: *arranjo e arranjo de documentos; conteúdo do documento e conteúdo dos documentos; autor e autoria; dependência contextual e dependência do contexto; ordem alfabética e ordenação alfabética* etc.) e identificar os termos sinônimos ou quase sinônimos (por exemplo: *KWIC* e *índice KWIC*; *radical e raiz* etc.) e eliminar como 'cabeça' de aglomerado aqueles termos ou expressões que se associam com um número muito pequeno de termos, normalmente inferior a 3 (por exemplo, *chave, código, comunicação direta, denotação, fala, indexador, interesse temático, nominação, regra* etc.), foram retidos, aproximadamente, 170 termos e expressões.

Convém esclarecer que a supressão desses termos ou expressões da lista de 'cabeças' de agrupamentos não significa de modo algum que não figurem na lista de termos associados a uma determinada 'cabeça' estatisticamente mais significativa. Assim, o termo *armazenamento*, que não se constitui em 'cabeça' de agrupamento, encontra-se entre os termos associados a *acesso*, o qual pode, com os critérios expostos, ser considerado como 'cabeça' de agrupamento. Da mesma forma, as expressões *descrição temática e lista de termos proibidos*, entre outras, que não são 'cabeças', encontram-se na lista de termos ou expressões integrantes de aglomerados referentes, respectivamente, à *catalogação e indexação automática*.

No anexo 1, encontra-se a relação de termos e expressões que poderiam ser considerados como possíveis 'cabeças' de agrupamentos, ordenados alfabeticamente, com indicação do número de termos que a eles se associam**.

* Esses termos, além de excessivamente genéricos, podem, em certos casos, por sua natureza polissêmica, conduzir a um aglomerado de termos associados que seria, na realidade, uma superposição de diversos aglomerados.

** Na relação de termos do anexo 1, não foram agrupados, num só, todos os termos sinônimos ou quase sinônimos (*autor* e a*utoria, KWOC e índice KWOC* etc.), já que os mesmos conceitos podem aparecer em formas diferentes nos diversos agrupamentos de termos.

TABELA 1

Fragmentos da tabela de freqüências dos termos considerados significativos (F_i igual ou maior que 2), no processo de indexação automática o Infodoc®.

TERMO	FREQ.	TERMO	FREQ.
AACR-2	2	Publicação	6
Acervo	6	Publicação periódica	2
Acesso	15	Qualidade de projeto	2
Acesso à informação	7	Radical	4
Acesso aleatório	2	Raiz	3
Acesso ao documento	4	Recuperação da informação	51
Acesso direto	3	Recuperação de dados	3
Acesso em linha	2	Recuperação de documento	3
Acesso seqüencial	2	Referência	6
		Referência bibliográfica	7
		Referência cruzada	7
Catalogação	8	Registro	19
Catalogação descritiva	2	Registro bibliográfico	11
Catálogo	17	Registro de informação	5
Catálogo alfabético de assunto	2	Registro documentário	3
Catálogo alfabético de autores	2	Registro informativo	2
Catálogo alfabético de títulos	2	.	
Catálogo coletivo	3	.	
Catálogo de assunto	2	Sigla	4
Catálogo de autor	2	Significação	8
Catálogo de títulos	2	Significação diferente	3
Catálogo dicionário	2	Significado	10
Catálogo ideográfico	2	Signo	8
Catálogo sistemático	3	Signo lingüístico	2
Categoria	5	Silêncio	2
Categoria fundamental	6	Símbolo	26
		Símbolo de classificação	2
		Sinal	3
Indexação	72	Sinônimo	5
Indexação automática	11	Sistema	2
Indexação controlada	2	Sistema conversacional	2
Indexação coordenada	3	Sistema de busca	3
Indexação hierárquica	21	Sistema de classificação	38
Indexação livre	2	Sistema de indexação	9
Indexação mecânica	2	Sistema de informação	13
Indexação pós-coordenada	4	Sistema de recuperação	5
Indexação pré-coordenada	3	Sistema de recuperação da informação	2
Indexação relacional	2	Sistema especialista	6
Indexador	2	Sistema informatizado	3
Índice	16	.	
Índice alfabético	5	.	
Índice alfabético de assuntos	2	Termo	67
Índice de classificação	9	Termo de busca	2
Índice KWIC	2	Termo de indexação	24
Índice KWOC	2	Termo específico	2
.		Termo genérico	4
.		Termo homônimo	4
Língua	5	Termo polissêmico	2
Linguagem	6	Termo preferencial	3
Linguagem artificial	5	Termo proibido	4
Linguagem de indexação	3	Tesauro	17
Linguagem documentária	15	Texto	11
Linguagem formal	2	Texto completo	3
Linguagem natural	5	Título	19
Lingüística	9	Título completo	3

TABELA 2

Fragmentos da tabela de freqüências dos termos considerados significativos (F_i e F_j igual ou maior que 2), com indicação das coocorrências dos pares associados (F_{ij}) e dos coeficientes de equivalência ou associação correspondente (E_{ij})

TERMO *i*	F_i	TERMO *j*	F_j	F_{ij}	E_{ij}
AACR-II	2	Acervo	6	1	0.O8
AACR-II	2	Catálogo	17	1	0.03
AACR-II	2	Descrição bibliográfica	23	2	0.O9
AACR-II	2	Elemento essencial	2	1	0.25
AACR-II	2	Entrada catalográfica	4	1	0.13
AACR-II	2	ISBD	14	2	0.14
AACR-II	2	Item	33	1	0.02
AACR-II	2	Norma de catalogação	2	2	1.00
.	.				
.	.				
Formato	2	Área de descrição física	2	1	0.25
Formato	2	Campo	6	1	0.08
Formato	2	Colação	4	1	0.13
Formato	2	Dado	38	1	0.01
Formato	2	Descrição bibliográfica	23	1	0.02
Formato	2	Descrição física	2	1	0.25
Formato	2	Documento	95	1	0.01
Formato	2	Ilustração	3	1	0.17
Formato	2	Inclusão de dados	2	1	0.25
Formato	2	Indexação	72	1	0.01
Formato	2	Informação	65	1	0.01
Formato	2	ISBD	14	1	0.04
Formato	2	Item	33	1	0.02
Formato	2	Representação codificada	2	1	0.25
Formato	2	Representação da informação	2	1	0.25
Formato	2	Zona	5	1	0.10
Formato	2	Zona de colação	2	1	0.25
Registro bibliográfico	11	Acesso	15	1	0.01
Registro bibliográfico	11	Arquivo seqüencial	2	1	0.03
Registro bibliográfico	11	Assunto	59	1	0.00
Registro bibliográfico	11	Base de dados	20	2	0.02
Registro bibliográfico	11	Busca bibliográfica	3	1	0.03
Registro bibliográfico	11	Campo	6	1	0.02
Registro bibliográfico	11	Catalogação	8	1	0.01
Registro bibliográfico	11	Catálogo coletivo	3	1	0.03
Registro bibliográfico	11	Classe	36	10	0.00
Registro bibliográfico	11	Classe geral	3	1	0.03
Registro bibliográfico	11	Classificação	54	1	0.00
Registro bibliográfico	11	Coleção	7	1	0.00
Registro bibliográfico	11	Dado	8	3	0.02
Registro bibliográfico	11	Descrição bibliográfica	23	1	0.00
Registro bibliográfico	11	Descritor	34	1	0.00
Registro bibliográfico	11	Documento	95	2	0.00
Registro bibliográfico	11	Entrada	20	1	0.00
Registro bibliográfico	11	Informação	65	1	0.00
Registro bibliográfico	11	Item	33	5	0.07
Registro bibliográfico	11	Item bibliográfico	7	2	0.00
Registro bibliográfico	11	Número do documento	2	1	0.05
Registro bibliográfico	11	Objeto	9	1	0.01
Registro bibliográfico	11	Ordenação alfabética	3	1	0.03
Registro bibliográfico	11	Ponto de acesso	3	1	0.03
Registro bibliográfico	11	Recuperação da informação	51	1	0.00
Registro bibliográfico	11	Recuperação do documento	4	1	0.02
Registro bibliográfico	11	Referência	6	1	0.02
Registro bibliográfico	11	Registro	19	1	0.00
Registro bibliográfico	11	Representação codificada	2	1	0.05
Registro bibliográfico	11	Ruído	2	1	0.05
Registro bibliográfico	11	Segmento de registro	2	1	0.05
Registro bibliográfico	11	Símbolo numérico	2	1	0.05
Registro bibliográfico	11	Sistema de busca	3	1	0.03
Registro bibliográfico	11	Sistema de classificação	35	1	0.00
Registro bibliográfico	11	Unidade de informação	5	2	0.07
Registro bibliográfico	11	Unidade documentária	3	1	0.03

APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Na impossibilidade de apresentar a totalidade dos agrupamentos de termos (com todas suas inter-relações) associados às 'cabeças' listadas no anexo 1, consideramos mais pertinente centrar a apresentação e discussão dos resultados em alguns exemplos significativos que mostram a potencialidade do método de análise da freqüência de coocorrência de pares de palavras-chave para 'mapear' uma determinada área de conhecimento e identificar associações conceituais do maior interesse para caracterizar os termos focais e o estado-da-arte da pesquisa científica nessa área, em um determinado momento e, por extensão, verificar sua evolução no tempo e no espaço.

No anexo 2, apresentam-se os aglomerados referentes a alguns termos e expressões associados a algumas 'cabeças'. São estas:

– *classificação;*

– *descrição bibliográfica;*

– *indexação;*

– *recuperação da informação;*

Como exemplo de subaglomerados, detalhados no anexo 3, foram escolhidos os seguintes:

– *catalogação* (incluído em *descrição bibliográfica*);

– *faceta* (incluído em *classificação*);

– *indexação automática* (incluído em *indexação*);

– *ISBD* (incluído em *descrição bibliográfica*);

– *Ranganathan* (incluído em *classificação e em faceta*);

– *tesauro* (incluído em *indexação*).

Nos anexos 2 e 3, são indicados os valores das freqüências de uso dos termos individuais *i* e *j* (*Fi* e *Fj*), os valores da freqüência de associação de cada par de termos (*Fij*) e o coeficiente de equivalência ou 'energia' de associação de cada par (*Eij*).

A partir do exame dos termos e expressões que figuram nas listas dos anexos 1 a 3, torna-se evidente o interesse de estudos deste tipo para identificar sinônimos e quase sinônimos, ou de termos relacionados semanticamente, os quais se agrupam naturalmente dentro dos aglomerados. Parece inútil insistir sobre a importância desses fatos na construção de tesauros ou na elaboração e manutenção de todo tipo de dicionários e léxicos que muito podem auxiliar, tanto no processo de indexação de documentos, quanto na busca e recuperação da informação. Como exemplo, pode-se mostrar as cadeias de termos a seguir:

Indexação automática - índice KWIC - KWIC - lista de termos proibidos - termos proibidos - índice KWOC - KWOC etc.

e

Truncagem - raiz - radical - desinência - sufixo - prefixo etc.

A título de comentário anedótico, oferecemos à apreciação do leitor duas associações encontradas (as quais, devido à sua baixa freqüência, não são evidentes nas listas ou nos aglomerados apresentados como exemplo neste trabalho), que, ao nosso ver, merecem um breve comentário. Trata-se da associação entre *número de chamada e estante*, e *Garfield* e *citação*, absolutamente lógicas, mas que não parece que tenham sido incluídas em nenhum tesauro ou outro tipo de vocabulário controlado de que temos conhecimento. Tais associações poderiam eventualmente facilitar uma pesquisa de informação em linguagem natural, em uma base de dados. Ainda no terreno anedótico, e falando em associações evidenciadas pelos agrupamentos, é interessante observar que, no estágio de avanço do *Glossário de Termos Técnicos de Ciência da Informação* [9], no momento em que foram selecionados os verbetes para compor o *corpus* deste trabalho, há quase dois anos, os verbetes referentes à *indexação automática* parece que focalizavam esta entrada somente sob o ângulo do índice *KWIC*, sem incluir ainda outros aspectos relevantes mais recentes.

Inútil também insistir sobre a importância de poder 'descobrir' relações entre termos e expressões que não são evidentes à primeira vista, dentro de uma visão rígida e estruturada hierarquicamente, como a que prevalece no desenvolvimento, manutenção e uso de tesauros e outros instrumentos terminológicos 'controlados', que impedem ver determinadas relações semânticas naturais que vão surgindo simultaneamente ao desenvolvimento de quaisquer áreas da ciência ou da tecnologia.

Uma análise mais aprofundada dos pares de termos associados e dos valores correspondentes de *Fij* e *Eij* permite maior aproximação do significado profundo dos aglomerados e das relações entre seus componentes. Em geral, *Eij* tende a apresentar valor mais elevado quando a abrangência do tema e a polissemia dos termos estão bem delimitados (ver, por exemplo, *Ranganathan* ou *faceta*, no anexo 3).

Dentro de um mesmo aglomerado, maiores valores de *Eij* representam, em geral, maior afinidade entre os termos integrantes do par (ver aglomerados nos anexos 2 e 3). O valor 0.00 para *Eij*, que aparece em alguns casos, não significa, como foi frisado anteriormente, que a coocorrência dos dois termos do par seja impossível (se assim for, eles não apareceriam nas listagens), mas que o valor do coeficiente de associação é inferior a 0.01, ou seja, que o par tem uma afinidade muito baixa.

Um termo que integra dois ou mais aglomerados diferentes pode apresentar valores bastante diferentes de *Eij*, para o par formado por ele e a 'cabeça' do aglomerado, indicando diferentes afinidades entre os respectivos pares. Assim, o termo *descritor*, que integra, entre outros, os aglomerados referentes à *indexação e recuperação da informação*, apresenta (ver anexo 2) nos pares correspondentes valores respectivos de 0.07 e 0.01, o que parece indicar que é mais provável utilizar o termo *descritor* em um contexto em que se fala de indexação do que em outro que trata de recuperação.

Por outra parte, quando um termo ou expressão pode ser encontrado em diferentes contextos, observa-se que o valor de *Eij*, para vários pares possíveis tende a diminuir ao aumentar a 'dispersão semântica'. Nesse caso, a 'força' de associação entre os pares de termos parece estar mais bem representada pelos valores de *Fij*, em geral igual ou maior que 3. Assim, nos aglomerados *indexação* e *recuperação da informação*, observa-se que, entre o grande número de pares de termos com *Eij* igual ou menor que 0.02, aqueles que possuem *Fij* igual ou maior que 3 são, em geral, mais fortemente relacionados.

A utilização dessas observações pode contribuir eficazmente para introduzir remissivas e referências cruzadas, na elaboração e manutenção de tesauros e vocabulários e, de modo especial, na elaboração e manutenção automáticas de dicionários de termos e expressões para indexação de textos com ajuda do computador e formulação de estratégias de busca em linguagem natural, para recuperação da informação.

Para visualizar a estrutura dos aglomerados e as relações entre seus elementos componentes, podem ser utilizados vários tipos de representação gráfica. Dentre eles, convém destacar a representação em rede (figura 1, a seguir). Nesse tipo de representação, é possível indicar a maior ou menor freqüência dos elementos componentes por círculos ou quadrados de tamanho proporcional aos valores das respectivas ocorrências (*Fi* e *Fj*), bem como destacar a 'força' de associação entre pares de termos (medida por *Eij* ou simplesmente por *Fij*), mediante linhas de enlace mais ou menos destacadas (por exemplo, linhas contínuas ou pontilhadas de espessuras diferentes).

Informações sobre a aplicação de outros tipos de representação mais complexos, que permitem distribuir, nos quatro quadrantes de um plano de coordenadas cartesianas, os diversos termos e expressões com indicação mais precisa da 'força' de ligação entre os aglomerados e de sua importância relativa ('centralidade' e 'densidade'), mediante os chamados 'diagramas estratégicos' que não se adequam ao propósito deste artigo, podem ser encontrados nos trabalhos de Whittaker[1], Callon, Courtial e Laville[7] e Cambrosio, Limoges, Courtial e Laville[13], já citados, assim como nas publicações de Courtial[19], Courtial e Law [20], Courtial, Callon e Sigogneau [21], Huot, Quoniam, e Dou [22], e Amudbavall e Raghavan [23] ***.

Neste trabalho, para não sobrecarregar nossa exposição, limitar-nos-emos a mostrar, a título de exemplo, a representação gráfica, em rede, do agrupamento referente a *Ranganathan* detalhado no anexo 3 (figura 2, a seguir). Observe-se que, como o tema está muito mais delimitado do que no caso, por exemplo, do aglomerado *classificação*, os valores de *Eij* são consideravelmente mais elevados (maior afinidade entre os pares de termos).

Para ilustrar o desdobramento dos agrupamentos em subagrupamentos cada vez menores, foram reunidos no anexo 4 os agrupamentos correspondentes a:

– *categoria fundamental;*

– *classificação dos dois pontos;*

– *energia;*

– *faceta fundamental;*

– *interesse temático;*

– *personalidade;*

– *PMEST;*

– *tempo.*

Todos eles relacionados com *faceta* e/ou *Ranganathan*. Os elementos integrantes do agrupamento faceta e dos subagrupamentos acima podem ser combinados com os itens da figura 2 de maneira a formar uma rede espacial de todos os termos inter-relacionados.

*** Centralidade (*centrality*) é a medida estatística da intensidade das ligações para um determinado agrupamento. Mede a coerência de um tópico e é representada pelo valor médio das ligações que existem entre as palavras-chave que integram o aglomerado. Uma forma de medi-la é calcular para cada aglomerado o valor médio de suas ligações internas.
Densidade (*density*) é a medida estatística da força das ligações que associam as palavras integrantes de um aglomerado. Caracteriza o papel desempenhado por um determinado tema no desenvolvimento global da área e é representado pelo valor médio das ligações entre um aglomerado e outro aglomerado vinculado ao primeiro por meio de algumas de suas palavras-chave ou, em uma definição mais técnica, a posição relativa de cada aglomerado dentro do 'mapa' global da área. Uma forma de calculá-la é somar os quadrados de todas ligações medidas pelo coeficiente de equivalência que o une a outros aglomerados.

CONCLUSÃO

O presente trabalho permitiu mostrar a potencialidade e interesse dos métodos de análise da coocorrência de palavras ou expressões significativas para 'mapear' um determinado campo do conhecimento, com sólidas bases teóricas e aplicações do maior interesse, em um leque de possibilidades que cobre da elaboração, estudo, manutenção e uso de instrumentos terminológicos os mais diversos, até a caracterização de uma área de pesquisa, o acompanhamento do desenvolvimento e evolução de um campo da ciência ou da tecnologia em um determinado período, ou, ainda, o estudo comparativo do estado-da-arte de um campo específico em várias instituições ou em momentos diferentes, assim como a realização de projeções sobre a evolução de uma área da ciência, como demonstrado em diversos trabalhos já referenciados [2-4, 5, 7, 12, 13, 14, 16-23].

Outras aplicações, tais como a definição ou avaliação da política de aquisição de documentos por parte de uma grande biblioteca universitária ou de um centro de documentação especializado, podem encontrar seu fundamento em estudos infométricos da análise da associação entre palavras, comparando os conteúdos temáticos dos documentos (livros, periódicos, atas de congressos, patentes etc.) com os programas curriculares, no primeiro caso, e com os programas de pesquisa, políticas institucionais e perfil dos usuários em ambos os casos, ou, ainda, analisando as citações usadas na produção científica dos especialistas ligados à organização.

Provavelmente, por não dizer com certeza, entre as áreas de pesquisa fundamental e aplicada que deverão conhecer uma grande expansão nos próximos anos, parecem encontrar-se justamente as aplicações da análise das associações de palavras. Com efeito, a indexação de documentos que devem incorporar-se ao 'magma informacional' da Internet (bases de dados, bibliotecas virtuais, documentos sobre um tema específico, identificação de especialistas, sejam estes pessoas ou instituí-

FIGURA 1
Representação em rede de um agrupamento de termos. Diferentes tipos de linha representam forças de enlace diferentes entre pares de termos. Diferentes tamanhos dos círculos que representam os termos representam diferentes valores de ocorrência.

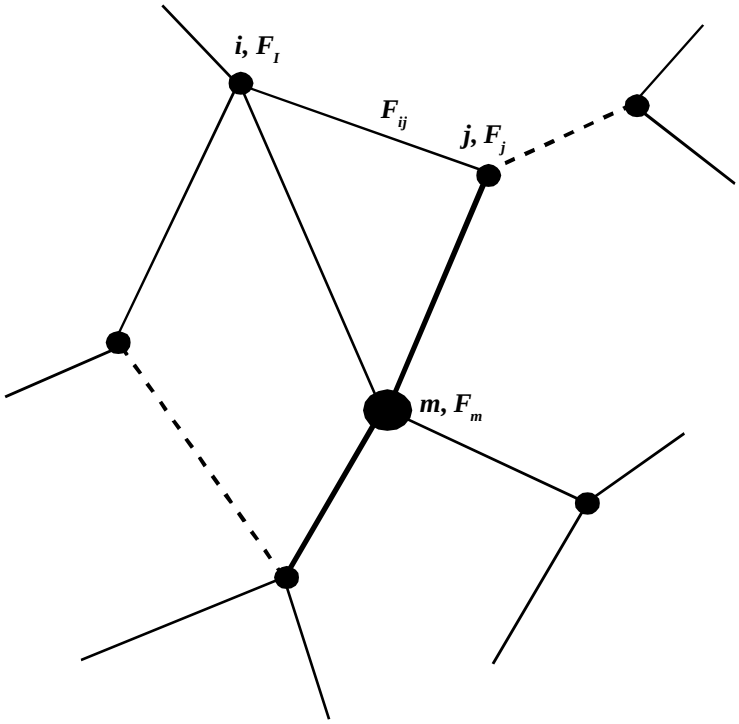
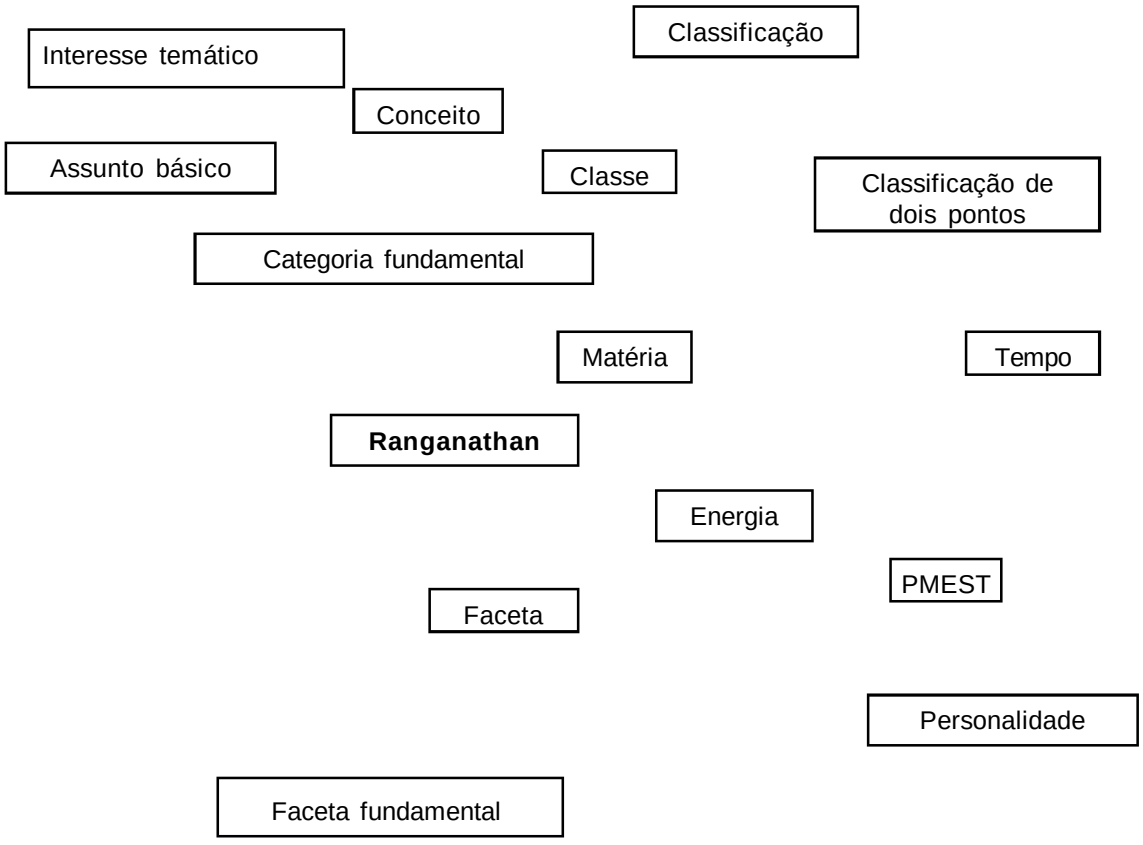


FIGURA 2
Representação do aglomerado referente a *Ranganathan*.



ções, notícias etc.), onde pela variedade de usuários e de fontes torna-se já absolutamente impossível pensar na consulta de tesauros rígidos (sempre desatualizados em, pelo menos, dois anos), exigirá a disponibilidade de novas ferramentas. Novos léxicos e dicionários, que se atualizem automaticamente com base em princípios infométricos solidamente fundamentados e que se incorporem aos motores de busca das bases de dados, deverão estar disponíveis imperativamente para serem incorporados aos sistemas, tanto na entrada como na recuperação, fazendo uma grande parte do trabalho que o usuário final ou o intermediário da informação têm de realizar ainda no momento atual, e isso com o risco permanente de deixar escapar grande quantidade da informação procurada.

A convergência da informática, da indexação automática e do desenvolvimento de motores de busca incorporados às bases de dados parece constituir a chave dos desenvolvimentos futuros da informação globalizada[24, 25].

REFERÊNCIAS BIBLIOGRÁFICAS

1. WHITTAKER, John. Creativity and Conformity in Science: Titles, keywords and Co-word Analysis. *Social Studies in Science*. v.19, 1989, p.473-496.

2. CALLON, Michel; LAW, John; Rip, Arie (eds). *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Basingstoke, Hants: MacMillan, 1986.

3. ROBREDO, Jaime. On Informetrics as a Tool for Forecasting. In: 5TH BIENNIAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS. River Forest. Il, 7-10 June 1995. *Proceedings*. Medford, NJ: Learned Information, 1995, p. 694. (Uma versão em português será publicada em breve.)

4. KING, J. A Review of Bibliometric and other Science Indicators and the Role in Research Evaluation. *Journal of Information Science*, v.13, 1987, p.261-276.

5. LAW, John; WHITTAKER, John. Mapping Acidification Research: A Test of the Co-word Method. *Scientometrics*, v.23, 1992, p.417-461.

6. LEYDESDORFF, L. The Search of Epistemic Networks. *Social Studies in Science*. v.21, n.1,1991, p.75-110.

7. CALLON, M.; COURTIAL, J.P.; LAVILLE, F. Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry. *Scientometrics*, v.22, n.1, 1991,p.155-205.

8. DIODATO, Virgil. *Dictionary of Bibliometrics*. New York; London: Norwood: Haworth, 1994. ISBN 1-56024-832-1.

9. CAVALCANTI, Cordélia R.; CUNHA, Murilo B. da. *Glossário de Termos Técnicos de Ciência da Informação*. Brasília DF: Universidade de Brasília. (Em preparação.)

10. ROBREDO, Jaime. Indexação automática de textos. *In*: 1° Encontro Nacional de Pesquisa em Ciência da Informação e Biblioteconomia. Belo Horizonte MG, 8-10 abr 1994. ANCIB. *Anais*. Campinas SP: ANCIB, 1994, p.15-17.

11. ROBREDO, Jaime. *InfoDoc: Manual do Usuário*. Brasília DF: Edição do autor, 1995. (Inclui disquete.)

12. LE MARC, M.; COURTIAL, J.P.; DROZDA SENKOVSKA, E.; PÉTARD, J.P.; Py, Y. The Dynamics of Research in the Psichology of Work from 1973 to 1987: From the Study of Companies to the Study of Professions. *Scientometrics*, v.21, n.1, 1991, p.60-68.

13. CAMBROSIO, A.; LIMOGES, C.; COURTIAL, J.P.; LAVILLE, F. Historical Scientometrics? Mapping over 70 Years of Biological Safety Research with Co-word Analysis. *Scientometrics*, v.27, n.2, 1993, p.119-143.

14. POLANCO, X. Scientometic Analysis of the Cognitive Sciences in Pascal. *INIST* Info, n.7, jul 1993.

15. POLANCO, X. Recherches sur les méthodes d'analyse stratégique de línformation scientifique e technique. *In*: Journée d'Étude sur les Systèmes d'Information Élaborés: Bibliométrie, Information Stratégique. Veille Technologique. Île Rousse. Société Française de Bibliométrie Appliquée. 5-7 Jun 1991. Tirage-à-part.

16. POLANCO, X. *et al*. À la recherche de la diversité perdue: est-il possible de mettre en évidence des élements hétérogènes d'un front de recherche? *Ibidem*. Tirage-à-part.

17. BASEVI, T.H.M.M. *Tendências na aplicação de formatos, sistemas cooperativos e redes de intercâmbio*: uma visão infométrica. Brasília DF: Universidade de Brasília/ Departamento de Ciência da Informação e Documentação, 1993. (Dissertação de mestrado.)

18. LIMA, A.C.C.C. *Sistemas especialistas aplicados à Ciência da Informação*: tendências para um futuro próximo baseadas em um estudo infométrico da literatura. Brasília DF: Universidade de Brasília/Departamento de Ciência da Informação e Documentação, 1993. (Dissertação de mestrado.)

19. COURTIAL, J.P. A Co-word Analysis of Scientometrics. *Scientometrics*, v.31, n.3, 1994, p.251-260.

20. COURTIAL, Jean-Paul.; LAW, John. A Co-Word Study of Artificial Intelligence. *Social Studies of Science*, v.19, 1989, p.301-311.

21. COURTIAL, J.P.; CALLON, M.; SIGOGNEAU, A. The use of Patents Titles for Identifying the Topics of Invention and Forecasting Trends. *Scientometrics*, v.26, n.2, 1993, p.231-242.

22. HUOT, Ch.; QUONIAM, L.; DOU, H. A. New Method for Analyzing Downloaded Data for Strategic Decision. *Scientometrics*, v.25, n.2, 1992, p.279-294.

23. AMUDBAVALLI, A.; RAGHAVAN, K.S. Co-word Analysis of Literature on Information Retrieval. *In*: 5TH BIENNIAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS. River Forest. Il, 7-10 June 1995. *Proceedings*. Medford, NJ: Learned Information, 1995, p.23-32.

24. ROBREDO, Jaime. Indexação e recuperação da informação na era das publicações virtuais. *In*: 4° SEMINÁRIO DE BIBLIOTECONOMIA E CIÊNCIA DA INFORMAÇÃO - GLOBALIZAÇÃO, INFORMAÇÃO E DESENVOLVIMENTO HUMANO SUSTENTÁVEL: um desafio para os profissionais da informação e da Comunicação. Goiânia GO, 1-4 jun 1997. Universidade federal de Goiás. *Anais*. (A ser publicado em breve.)

25. ROBREDO, Jaime. Indexação automática e infometria: um casamento que está dando certo. *In*: 18° CONGRESSO BRASILEIRO DE BIBLIOTECONOMIAE DOCUMENTAÇÃO. São Luís MA, 20-24 jul 1997. *Anais eletrônicos*.

The use of informetrics for identifying the scope of the basic terminology related to indexing and retrieval

Abstract

Co-word analysis offers the possibility of statiscally measuring the associative strength between pairs of keywords and, by using the values found, of mapping the dynamics of a scientific field in a given moment. The identification of clusters of keywords and the analysis of the strength of the links between pairs of keywords in the clusters show the way for important applications, ranging from the building up of special lexicons, to the development of logical tools for optimyzing automatic indexing and retrieval processes, as well as the mapping of the evolution of interest on key topics in scientific research. An application of co-word analysis to identify the scope of the basic terminology related to indexing and retrieval is described.

Keywords

Basic vocabulary; Indexing; Information retrieval; Infometrical techniques.

Artigo aceito para publicação em 17-2-98.

Jaime Robredo

Pesquisador associado senior.
Departamento de Ciência da Informação e Documentação.
Universidade de Brasília.

E-mail: jrobredo@brnet.com.br

Murilo Bastos da Cunha

Professor titular.
Departamento de Ciência da Informação e Documentação.
Universidade de Brasília.

E-mail: murilobc@guarany.unb.br

ANEXO 1

Relação dos termos considerados "cabeça" de agrupamento

TERMO	FREQÜÊNCIA	TOTAL TERMOS ASSOC.
AACR-II	2	3
Acervo	6	6
Acesso	15	32
Acesso a informação	7	5
Acesso ao documento	4	4
Análise documentária	5	6
Área de conhecimento	4	2
Armazenamento de informação	3	6
Arquivo	16	19
Arquivologia	4	6
Arranjo	5	4
Assunto	59	57
Assunto básico	5	7
Assunto composto	3	5
Autor	17	17
Autoria	5	5
Banco de dados	11	20
Base de dados	20	23
Bibliografia	11	15
Bibliometria	2	3
Biblioteca	12	23
Busca	7	9
Busca da informação	11	14
Cabeçalho de assunto	13	17
Campo	6	8
Catalogação	8	7
Catálogo	17	32
Catálogo alfabético de autores	2	5
Catálogo sistemático	3	10
Categoria fundamental	6	10
Citação	8	10
Citação bibliográfica	4	4
Classe	36	35
Classe multidimensional	2	4
Classificação	54	56
Classificação bibliográfica	7	7
Classificação de segurança	2	3
Classificação de dois pontos	5	4
Classificação facetada	3	3
Classificação unidimensional	2	6
Coerência da indexação	2	3
Colação	4	5
Coleção	7	7
Computador	7	7
Comunicação	5	8
Conceito	30	28
Conhecimento	6	4
Consistência na indexação	3	4
Consulta	5	10
Conteúdo temático	13	16
Contexto	5	4
Dado	38	40
Data	2	7
Data de publicação	4	8
Dependência contextual	2	4
Dependência do contexto	2	4
Descrição bibliográfica	23	15
Descritor	34	30
Documento	95	102
Edição	4	6
Editor	2	8
Entrada	20	24
Entrada bibliográfica	3	3
Entrada de assunto	5	6
Esquema de classificação	23	14
Estratégia de busca	8	5

Expressão	8	9
Extensão	4	4
Faceta	17	22
Faceta fundamental	3	5
Ficha	2	8
Fichário	6	5
Garfield	2	3
Gênero	8	5
Homônimo	4	3
Idéia	7	3
Imprensa	3	7
Iindexação	17	86
Indexação automática	11	11
Indexação coordenada	3	8
Indexação pós-coordenada	4	9
Indexação pré-coordenada	3	9
Índice	16	28
Índice alfabético	5	3
Índice de classificação	9	8
Índice KWOC	2	3
Informação	65	70
Instrumento de busca	3	5
ISBD	14	22
Item	33	29
Item bibliográfico	7	4
Item recuperado	3	3
KWIC	4	8
KWOC	4	6
Língua	5	4
Linguagem	6	3
Linguagem artificial	5	3
Linguagem documentária	15	17
Lingüística	9	5
Lista	3	3
Livro	4	3
Lógica	9	9
Lógica tradicional	3	5
Lugar de publicação	2	6
Matéria	4	4
Memória	2	7
Método de Indexação	6	10
Multidimensionalidade	2	5
Nome	4	5
Norma de catalogação	2	3
Notação	12	4
Objeto	9	4
Operador booleano	5	4
Ordem alfabética	11	11
Ordenação de catálogo	2	3
Página de rosto	5	3
Palavra	57	51
Palavra-chave	12	13
Palavra significativa	4	7
Permutação de palavras	2	3
Personalidade	2	4
Pesquisa documentária	3	3
PMEST	3	6
Polissemia	2	3
Ponto de acesso	3	3
Pós-coordenação	3	7
Pré-coordenação	3	6
PRECIS	4	6
Precisão	4	6
Predicável	4	5
Publicação	6	7
Publicação periódica	2	3
Radical	4	3
Raiz	3	2
Ranganathan	11	14
Recuperação da informação	51	58
Recuperação de documento	4	3

Referência	6	5
Referência bibliográfica	7	4
Referência cruzada	7	3
Registro	19	26
Registro bibliográfico	11	6
Registro de informação	5	12
Relação entre termos	5	3
Relação semântica	3	6
Remissiva	5	5
Representação automática	8	13
Responsabilidade	4	6
Resumo	8	6
Revogação	5	5
Sentido	11	11
Série	8	10
Significação	8	5
Significação diferente	3	4
Significado	10	7
Signo	8	6
Símbolo	26	22
Sintaxe	5	8
Sistema	2	8
Sistema de classificação	38	33
Sistema de indexação	9	8
Sistema de informação	13	12
Sistema de recuperação	5	4
Sistema de recuperação da informação	2	3
Subclasse	5	6
Subdivisão de assunto	5	3
Subdivisão de classe	4	5
Tema	3	4
Tempo	3	5
Termo	62	68
Termo de busca	2	3
Termo de indexação	24	23
Termo genérico	4	6
Termo homônimo	4	4
Termo polissêmico	2	4
Termo preferencial	3	3
Termo proibido	4	5
Tesauro	17	15
Texto	11	11
Título	19	28
Unidade de informação	5	4
Unidimensionalidade	4	9
Unitermo	4	9
Usuário	8	10
Vocabulário	5	7
Zona	8	5

ANEXO 2

Exemplos de aglomerados de termos significativos, com indicação do número de coocorrências (F_{ij}) e do valor do coeficiente de associação (E_{ij}). Os valores entre parêntese indicam as freqüências do termo "cabeça" de aglomerado e dos termos associados.

'CABEÇA' DE AGRUPAMENTO	TERMOS ASSOCIADOS	F_{ij}	E_{ij}
Classificação (54)	Arquivo (16)	2	0.00
	Assunto (59)	15	0.07
	Autor (17)	2	0.02
	Banco de dados (11)	2	0.01
	Biblioteca (12)	2	0.01
	Catalogação (8)	3	0.02
	Categoria fundamental (6)	4	0.05
	Citação (6)	2	0.01
	Classe (36)	14	0.10
	Classificação bibliográfica (7)	2	0.00
	Classificação dos dois pontos (5)	3	0.03
	Conceito (30)	6	0.02
	Conhecimento (6)	3	0.03
	Conteúdo temático (13)	2	0.01
	Descrição temática (2)	1	0.01
	Elemento (6)	3	0.03
	Esquema de classificação (23)	13	0.14
	Faceta (17)	5	0.03
	Faceta fundamental (3)	2	0.02
	Indexação (72)	7	0.01
	Índice (16)	3	0.01
	Notação (12)	5	0.04
	Ranganathan (11)	5	0.04
	Recuperação da informação (51)	3	0.00
	Registro (19)	3	0.01
	Série (8)	3	0.02
	Símbolo (26)	9	0.06
	Sistema de classificação (36)	16	0.12
	Sublcasse (5)	2	0.01
	Tabela de classificação (2)	2	0.04
	Tema (3)	2	0.02
	Tempo (3)	2	0.02
	Título (19)	2	0.00
Descrição bibliográfica (23)	AACR-II (2)	2	0.09
	Acervo (6)	3	0.07
	Área de edição (2)	2	0.09
	Área de publicação (3)	2	0.06
	Área específica de material e tit. (2)	2	0.09
	Autoria (5)	2	0.03
	Biblioteca (12)	2	0.01
	Campo (6)	3	0.05
	Catalogação (8)	3	0.05
	Catálogo (17)	3	0.02
	Colação (4)	2	0.04
	ISBD (14)	12	0.44
	Responsabilidade (4)	3	0.10
	Zona (5)	4	0.14
	Zona de colação (2)	2	0.09
Indexação (72)	Assunto (59)	17	0.07
	Autor (17)	3	0.01
	Base de dados (20)	3	0.01
	Busca de informação (11)	3	0.01
	Cabeçalho de assunto (13)	3	0.01
	Catálogo (17)	3	0.01
	Citação (8)	3	0.02
	Classificação (54)	7	0.01
	Conceito (30)	6	0.02
	Consistência na indexação (3)	3	0.04
	Conteúdo temático (13)	9	0.09
	Descritor (34)	13	0.07
	Entrada (20)	4	0.01
	Indexação automática (11)	5	0.03

	Indexação coordenada (31)	3	0.04
	Indexação pós-coordenada (4)	4	0.06
	Indexação pré-coordenada (3)	3	0.04
	Índice (16)	7	0.07
	Linguagem de indexação (3)	3	0.04
	Linguagem documentária (15)	7	0.05
	Método de indexação (6)	6	0.08
	Palavra (57)	12	0.04
	Palavra-chave (12)	6	0.04
	Pós-coordenada (3)	3	0.04
	Pré-coordenada (3)	3	0.04
	PRECIS (4)	4	0.04
	Recuperação da informação (51)	12	0.04
	Relação entre termos (5)	3	0.03
	Representação temática (8)	4	0.03
	Significado (10)	3	0.01
	Símbolo (26)	4	0.01
	Sintaxe (5)	3	0.03
	Sistema de classificação (38)	3	0.00
	Sistema de indexação (9)	9	0.13
	Sistema de informação (13)	2	0.00
	Termo (67)	31	0.02
	Termo de indexação (24)	23	0.31
	Termo proibido (4)	2	0.01
	Tesauro (17)	5	0.02
	Texto (11)	2	0.01
	Título (19)	5	0.02
	Uniformidade na indexação (2)	2	0.03
	Unitermo (4)	3	0.03
	Vocabulário (5)	2	0.01
	Vocabulário controlado (3)	3	0.03
Recuperação da informação (51)	Arquivo (16)	4	0.02
	Assunto (59)	8	0.02
	Banco de dados (11)	5	0.04
	Base de dados (20)	7	0.05
	Busca (7)	3	0.03
	Busca da informação (11)	8	0.11
	Classificação (54)	3	0.00
	Coleção (7)	3	0.03
	Conceito (30)	3	0.01
	Contexto temático (13)	4	0.02
	Descritor (34)	3	0.01
	Documento recuperado (2)	2	0.04
	Estratégia de busca (8)	6	0.09
	Formulação da pergunta (2)	2	0.04
	Indexação (72)	12	0.04
	Índice (16)	4	0.02
	Item recuperado (3)	3	0.06
	Linguagem documentária (15)	4	0.00
	Precisão (4)	3	0.04
	Revogação (5)	3	0.04
	Símbolo (26)	3	0.01
	Sistema de informação (13)	4	0.02
	Ssistema de recuperação (5)	5	0.10
	Sistema de recuperação da informação (2)	2	0.04
	Termo de indexação (24)	4	0.0
	Vocabulário controlado (3)	3	0.06

NOTA: não foram incluídos os termos dado, documento, informação, item, palavra, registro, termo por serem excessivamente genéricos.

ANEXO 3

Exemplos de subaglomerados de termos significativos, com indicação do número de co-ocorrências (F_{ij}) e do valor do coeficiente de associação (E_{ij}). Os valores entre parêntese indicam as freqüências do termo 'cabeça' de subaglomerado e dos termos associados.

'CABEÇA' DE AGRUPAMENTO	TERMOS ASSOCIADOS	F_{ij}	E_{ij}
Catalogação (8)	Assunto (59)	2	0.01
	Catálogo (17)	3	0.27
	Classificação (54)	3	0.02
	Descrição bibliográfica (23)	3	0.05
	Descrição temática (2)	2	0,25
	Indexação (72)	2	0.01
	Referência (6)	2	0.08
Faceta (17)	Assunto (59)	3	0.01
	Assunto básico (5)	2	0.05
	Assunto composto (3)	2	0.08
	Categoria fundamental (6)	2	0.04
	Classe multidimensional (2)	2	0.12
	Classe unidimensional (2)	2	0.12
	Classificação (54)	5	0.03
	Classificação dos dois pontos (5)	2	0.05
	Classificação facetada (3)	3	0.18
	Classificação unidimensional (2)	2	0.12
	Divisão de uma faceta (2)	2	0.12
	Foco (3)	2	0.08
	Isolado (2)	2	0.12
	Multidimensionalidade (2)	2	0.12
	Personalidade (2)	2	0.12
	Ranganathan (4)	3	0.05
	Relação semântica (3)	2	0.08
	Sistema de classificação (38)	3	0.04
	Subclasse (5)	4	0.19
	Subdivisão de classe (4)	3	0.13
	Termo (67)	2	0.00
	Unidimensionalidade (4)	4	0.24
Indexação automática (11)	Indexação (75)	5	0.03
	Indexação mecanizada (2)	2	0.18
	Índice (16)	2	0.02
	KWIC (4)	3	0.20
	KWOC (4)	2	0.01
	Lista de termos proibidos (2)	2	0.18
	Palavra (57)	5	0.04
	Permutação de palavra (2)	2	0.18
	Termo (67)	3	0.01
	Termo de indexação (24)	2	0.02
	Título (19)	4	0.08
ISBD (8)	AACR-II (2)	2	0.14
	Área de descrição física (2)	2	0.14
	Área de distribuição (2)	2	0.14
	Área de edição (2)	2	0.14
	Área de publicação (2)	3	0,10
	Área de publicação, distribuição (3)	2	0.14
	Área específica de material e tit. (2)	2	0.14
	Autoria (5)	2	0.06
	Colação (4)	2	0.07
	Data (2)	2	0.14
	Data de publicação (4)	2	0.07
	Descrição bibliográfica (23)	11	0.45
	Distribuição (2)	2	0.14
	Distribuidor (2)	2	0.14
	Edição (4)	2	0.07
	Editor (2)	2	0.14
	Imprensa (3)	2	0.10
	ISSN (2)	2	0.14
	Lugar de publicação (2)	2	0.14
	Norma de catalogação (2)	2	0.14
	Responsabilidade (4)	3	0.16

	Zona (5)	5	0.36
	Zona de colação (2)	2	0.14
Ranganathan (11)	Assunto básico (5)	2	0.07
	Categoria fundamental (6)	6	0.55
	Classe (36)	6	0.01
	Classificação (54)	5	0.04
	Classificação dos dois pontos (5)	5	0.45
	Conceito (30)	2	0.01
	Energia (3)	2	0.12
	Faceta (17)	3	0.05
	Faceta fundamental (3)	3	0.27
	Interesse temático (2)	2	0.18
	Matéria (4)	2	0.09
	Personalidade (2)	2	0.18
	PMEST (3)	3	0.27
	Tempo (3)	2	0.12
Tesauro (17)	Assunto (59)	4	0.02
	Conceito (30)	5	0.05
	Descritor (34)	12	0.25
	Elaboração de tesauro (2)	2	0.12
	Expressão (8)	3	0.01
	Indexação (72)	5	0.02
	Linguagem documentária (15)	3	0.04
	Termo (17)	8	0.06
	Termo genérico (4)	2	0.06
	Termo referencial (3)	2	0.08
	Termo proibido (4)	2	0.06
	Vocabulário (5)	2	0.05

NOTA: não foi incluído o termo documento por ser excessivamente genérico.

ANEXO 4

Exemplos de subaglomerados de termos significativos referentes aos aglomerados *Faceta* e *Ranganathan* (mostrados no anexo 3), com indicação do número de coocorrências (F_{ij}) e do valor do coeficiente de associação (E_{ij}). Os valores entre parêntese indicam as freqüências do termo 'cabeça' de subaglomerado e dos termos associados.

'CABEÇA' DE AGRUPAMENTO	TERMOS ASSOCIADOS	F_{ij}	E_{ij}
Categoria fundamental (6)	Classificação (54)	4	0.05
	Classificação dos dois pontos (5)	4	0.53
	Conceito (30)	2	0.02
	Energia (3)	2	0.22
	Faceta (17)	2	0.04
	Faceta fundamental (3)	3	0.50
	Matéria (4)	2	0.17
	Personalidade (2)	2	0.33
	PMEST (3)	3	0.50
	Ranganathan (11)	6	0.55
Classificação dos dois pontos (5)	Classificação (54)	3	0.03
	Conceito (30)	2	0.03
	Ranganathan (11)	5	0.45
Energia (3)	Categoria fundamental (6)	2	0.02
	Classificação (54)	2	0.02
	Ranganathan (11)	2	0.12
FACETA FUNDAMENTAL (3)	Categoria fundamental (6)	3	0.50
	Classificação (54)	2	0.02
	Classificação dos dois pontos (5)	2	0.27
	PMEST (3)	2	0.44
	Ranganathan (11)	3	0.27
INTERESSE TEMÁTICO (2)	Ranganathan (11)	2	0.18
PERSONALIDADE (2)	Categoria fundamental (6)	2	0.33
	Classificação (54)	2	0.04
	Faceta (17)	2	0.12
	Ranganathan (11)	2	0.18
PMEST (3)	Categoria fundamental (6)	3	0.50
	Classificação (54)	2	0.02
	Faceta fundamental (3)	2	0.44
	Matéria (4)	2	0.33
	Ranganathan (11)	3	0.27
	Tempo (3)	2	0.44
TEMPO (3)	Assunto (59)	2	0.22
	Categoria fundamental (6)	2	0.22
	Classificação (54)	2	0.22
	PMEST (3)	2	0.44
	Ranganathan (11)	2	0.12