COMUNICAÇÕES

MEDIDAS DE CONSISTÊNCIA DA INDEXAÇÃO; INTERCONSISTÊNCIA*

Lena Vania Ribeiro Pinheiro
Universidade Federal do Pará

1 INTRODUÇÃO

Na recuperação da informação a indexação é um processo fundamental e deve ser, tanto quanto possível, consistente. A consistência da indexação reflete similaridades ou diferenças de termos de indexação, isto é, diferentes reações de indexadores processando a informação.

O objetivo do presente estudo é medir a consistência de um grupo de indexadores, através do grau de concordância ou discordância na escolha de um termo ou de um conjunto de termos para a representação do conteúdo de um documento.

Indexar é um processo intelectual altamente subjetivo, portanto, é praticamente impossível uma consistência absoluta. Os termos de indexação denotam conceitos e conotam classes. Os índices funcionam como uma ligação entre a fonte de informação e o usuário. Para Rothman "o valor do índice aumenta na razão direta do tamanho, diversidade e complexidade da fonte e deve minimizar os esforços da pesquisa". O indexador deve se imbuir do papel do usuário para ser capaz de incluir os termos relevantes e excluir os irrelevantes.

A indexação envolve julgamento e, conseqüentemente, oscila muito no seu nível de concordância e apresenta discrepâncias.

Estudos de consistência, em diferentes áreas, realizados por Hooper, Slamecka, Jacoby, Painter, Rodgers, Korotkin, Oliver, Schultz, Orr e outros, mostram resultados com variações entre 10 e 80%. A consistência depende das condições de desempenho da indexação, da experiência dos indexadores e de instrumentos de ajuda à indexação, tais como regras em manuais, vocabulários controlados etc. Sendo um exercício de seleção e de decisão, envolve também lógica e intuição. A consistência é consideravelmente aumentada quando são adotadas tabelas de classificação, thesauri ou restrições no uso dos termos.

Na medida da qualidade de indexação podem ser adotados pesos para os termos pois, para medir a consistência, o problema é a suposição de uma igualdade de relevância para os termos selecionados, o que não é exato.

Não existem critérios objetivos de indexação e os termos podem ser considerados mais pertinentes, mais informativos, mais relevantes etc. As medidas de consistência pretendem eliminar, ao menos parcialmente, os defeitos de outras medidas e mesmo as falhas de indexação.

Medidas de consistência podem expressar, além da concordância de indexadores quanto aos termos, a significância, desses termos.

Entre alguns conceitos de indexação destaca-se o de Maron e Kuhns, inicialmente conhecido como indexação automática, depois indexação probabilística e estatística, na qual há uma pertinência gradativa na ordenação dos termos, que conduz à interação de dois conjuntos de termos pela existência de associações. Os graus ou pesos possibilitam caracterizar melhor o conteúdo da informação do documento, automaticamente. Na aritmética da indexação, o número de relevância é a medida de relevância provável de um documento para o usuário.

Segundo Schneider, a coordenação de termos pode ser feita pela classificação hierárquica do assunto, e a classificação enumerativa seria a melhor solução para o problema de informação.

Tinker realizou os estudos mais importantes em indexação, considerando o fator humano e derivando uma fórmula para consistência cujo cálculo é a divisão dos termos em concordância $A \cap B$, pelo total de termos únicos (não concordantes) $A \cup B$.

Para medir a consistência de um grupo de indexadores são aplicadas fórmulas para cada par no conjunto, ou melhor, do indexador A para o B e vice-versa:

$$C = \frac{A \cap B}{A}$$

$$C = \frac{A \cap B}{B} \cdot A \cap B$$

correspondem aos termos comuns aos indexadores A e B. A, o total de termos do indexador A, e B, o total do indexador B.

Há dois tipos de testes: de interconsistência e de intraconsistência. No primeiro é medida a consistência de indexação entre dois indexadores ou um grupo de indexadores. Na intraconsistência é testada a consistência do indexador em relação a si mesmo através do tempo, isto é, em diferentes fases.

Nas pesquisas realizadas sobre indexação concluiu-se que, quando os termos são extraídos de títulos, sub-títulos e resumos, a consistência é maior.

A indexação pode ser uma representação derivativa, se parte de um conhecimento do qual é criado o vocabulário e pode ser extraída do documento.

* *Comunicação baseada em trabalho apresentado à disciplina Indexação e Thesaurus do Curso de Pós-Graduação (Mestrado) em Ciência da Informação do IBICT/UFRJ, ministrada pelos professores Gilda Maria Braga e Tefko Saracevic.*

quando utiliza a linguagem do próprio documento. A indexação através do texto do documento apresenta maior inconsistência pois a fonte é mais ampla e há maior probabilidade de termos para seleção.

Zunde e Dexter, partindo de resultados experimentais quanto à exaustividade de indexação, observaram que o aumento de informação como resultado de um grupo de indexadores obedece às mesmas leis de dispersão da informação (Bradford), de produtividade científica (Lotka) e de distribuição de vocabulário (Zipf).

2 MATERIAL

O artigo indexado é de autoria de Jean-Claude Gardin, cujo título é "Document Analysis and Linguistic Theory". Foi publicado no The Journal of Documentation, volume 29, número 2, em junho de 1973. Possui trinta e duas páginas (p. 137-169)

O grupo de indexadores é constituído por dez (10) mestrandos do Curso de Pós-Graduação (Mestrado) do Instituto Brasileiro de Informação em Ciência e Tecnologia — IBICT, sendo oito (8) bibliotecários e dois de outras áreas (História e Letras).

3 MÉTODO

A indexação do artigo foi livre, permitindo alta exaustividade. Não foram feitas quaisquer restrições aos indexadores nem tampouco lhes foram fornecidos instrumentos de ajuda à indexação, tais como manuais ou thesauri.

A indexação foi realizada no espaço de uma hora, dividida em tempos de dez (10) minutos cada. O artigo foi lido durante os primeiros dez minutos e anotados os termos, assim sendo repetido, de dez em dez minutos, até completar uma hora.

Os termos selecionados por cada indexador foram listados, dentro dos tempos correspondentes. Foram comparados os termos de cada um dos indexadores, em relação aos outros nove (9).

Para medida de consistência recíproca foi utilizada a fórmula

$$C = \frac{A \cap B}{A \cup B}$$

Para medir a consistência de um indexador em relação a outro e vice-versa, foram adotadas as fórmulas

$$C = \frac{A \cap B}{A} \quad e \quad C = \frac{A \cap B}{B}$$

Foram considerados termos coincidentes, ou comuns, os que apareceram no singular e plural. Ex: método interpretativo e métodos interpretativos, thesaurus e thesauri, índice e índices etc. Os termos compostos ligados por preposição foram também assim contados: análise de conteúdo e análise do conteúdo, análise de documento e análise do documento. Todos os outros casos de pequenas diferenças na escolha dos termos de indexação não foram incluídos entre os termos comuns, tais como, cálculo relacionado e cálculo relacional, análise do documento e análise documentária, instrumentos semânticos e instrumentos da semântica, e outros.

4 RESULTADOS

4.1 Número de termos em relação aos tempos

Os termos selecionados pelos indexadores a cada um dos seis tempos, de dez minutos, apresentam o seguinte resultado:

Indexadores	Número de termos						
	1º tempo	2º t.	3º t.	4º t	5º t.	6º t.	Total
A	13	13	14	9	8	5	62
B	17	22	9	0	0	0	48
C	16	7	8	19	5	0	55
D	27	15	17	3	0	0	62
E	23	11	10	4	0	0	48
F	16	5	5	5	4	2	37
G	20	9	11	10	5	0	55
H	9	2	5	4	0	0	20
I	11	9	8	2	1	0	31
J	32	19	12	0	0	0	63
Total	**184**	**112**	**99**	**56**	**23**	**7**	**481**

Nota-se que, para um total de quatrocentos e oitenta e um (481) termos, nos primeiros três tempos, isto é, trinta (30) minutos, foram indexados trezentos e noventa e cinco (395) termos, ou melhor, 82% do total. Para dois indexadores (B e J) não foram necessários os três últimos tempos pois esgotaram a indexação dos itens na primeira meia hora. Três indexadores (D, E e H) não selecionaram termos no quinto e sexto tempos. Apenas dois indexadores incluíram itens nos últimos dez minutos (A e F).

Houve coincidência no total de itens indexados entre os indexadores A e D (62 termos), B e E (48 termos) e C e G (55 termos). Embora nem sempre o primeiro tempo contenha o maior número de termos, há uma tendência à diminuição de termos nos tempos subseqüentes, sobretudo a partir do quarto tempo.

Não houver economia na indexação, a julgar pela quantidade de termos indexados, mesmo considerando a extensão do texto do artigo, trinta e duas (32) páginas.

4.2 Consistência

A medida de consistência do grupo de indexadores entre si, aos pares, é demonstrada no quadro abaixo:

Indexadores	Indexadores									
	A	B	C	D	E	F	G	H	I	J
A		0,30	0,23	0,20	0,23	0,06	0,13	0,17	0,20	0,21
B			0,24	0,22	0,26	0,08	0,14	0,17	0,25	0,24
C				0,19	0,24	0,05	0,12	0,15	0,17	0,20
D					0,18	0,08	0,15	0,12	0,16	0,09
E						0,06	0,15	0,13	0,16	0,12
F							0,10	0,11	0,07	0,06
G								0,19	0,11	0,09
H									0,24	0,16
I										0,17
J										

A consistência maior ocorreu entre os indexadores A e B (0,30), seguida por 0,26, entre os indexadores B e E, depois 0,25, B e I; 0,24, A e C, B e J, C e E e H e I; 0,23, consistência entre os indexadores A e C e A e E e 0,22 (B e D). A consistência mais baixa foi de 0,05 (indexadores C e F), seguida de 0,06 (A e F, B e F, E e F, F e J), de 0,07 (F e I), 0,08 (D e F) e 0,09 (D e J, G e J). Observa-se que a baixa consistência envolve principalmente o indexador F, seguindo pelo indexador J. A média de consistência no conjunto de indexadores pode ser considerada baixa, 0,15.

A consistência foi medida por pares de indexadores, nos dois sentidos, isto é, de A para B e de B para A e assim sucessivamente, seguindo as fórmulas:

$$C = \frac{A \cap B}{A} \quad e \quad C = \frac{A \cap B}{B}$$

Indexadores	Indexadores										T	M
	A	B	C	D	E	F	G	H	I	J		
A	1	0,41	0,35	0,33	0,33	0,09	0,22	0,19	0,25	0,35	3,52	0,35
B	0,54	1	0,41	0,41	0,41	0,10	0,27	0,20	0,33	0,45	4,12	0,41
C	0,4	0,36	1	0,34	0,36	0,09	0,21	0,18	0,23	0,36	3,53	0,35
D	0,33	0,32	0,30	1	0,27	0,12	0,25	0,14	0,20	0,17	3,10	0,31
E	0,43	0,41	0,41	0,35	1	0,10	0,29	0,16	0,22	0,25	3,62	0,36
F	0,16	0,13	0,13	0,21	0,13	1	0,24	0,16	0,13	0,16	2,45	0,24
G	0,25	0,23	0,21	0,29	0,25	0,16	1	0,21	0,16	0,18	2,94	0,29
H	0,6	0,5	0,8	0,45	0,4	0,3	0,6	1	0,5	0,55	5,40	0,54
I	0,5	0,51	0,41	0,41	0,35	0,16	0,29	0,32	1	0,45	4,40	0,44
J	0,34	0,34	0,31	0,17	0,19	0,09	0,15	0,17	0,22	1	2,98	0,29
Total	4,21	4,21	4,03	2,96	3,69	2,21	3,52	2,73	3,24	3,92		
Média	0,42	0,42	0,40	0,29	0,36	0,22	0,35	0,27	0,32	0,39		

A medida de consistência foi igual para os indexadores A e D, B e E, C e G e D e J.

A média de consistência de cada indexador em relação aos demais é mostrada nos dois sentidos A → B e B → A. A mínima foi de 0,22 e, a máxima, 0,54. As médias variaram muito e apenas um indexador (E) apresentou médias coincidentes tanto na sua consistência quanto aos outros indexadores, como vice-versa.

4.3 Freqüência de termos

Dos termos indexados, um pelo menos apresenta freqüência absoluta (metalinguagem), seguido por análise de documento, ciência da informação, linguagem natural, análise de conteúdo e lingüística, não indexados por apenas um indexador e, portanto, freqüência 9. O quadro abaixo expressa a freqüência em ordem decrescente até freqüência 5, em vinte e sete (27) termos. Não foram levantadas as freqüências mais baixas porque o estudo mede consistência de indexadores e não de termos. A freqüência serve para uma análise comparativa com as medidas de consistência calculadas.

Nota-se que os termos indexados são, na sua maioria, termos simples (uma palavra) ou compostos por duas e, no máximo, por três palavras.

Os termos, quando não incluídos na seleção dos indexadores, serão identificados pelo espaço em branco, e a inclusão pela letra x.

Termos	Indexadores										Total/freqüência
	A	B	C	D	E	F	G	H	I	J	
Metalinguagem	x	x	x	x	x	x	x	x	x	x	10
Análise do documento	x	x	x	x	x		x	x	x	x	9
Ciência da Informação	x	x	x	x	x	x	x	x	x		9
Linguagem natural	x	x	x	x	x		x	x	x	x	9
Análise de conteúdo	x	x	x	x	x		x	x	x	x	9
Lingüística	x	x	x	x	x		x	x	x	x	9
Índices	x	x	x	x	x			x	x	x	8
SYNTOL	x	x		x	x	x	x		x	x	8
Teoria lingüística	x	x	x	x	x				x	x	7
Métodos de tabulação	x	x			x		x	x	x	x	7
Concordância	x	x	x	x	x				x	x	7
DEACON	x	x	x	x	x				x	x	7
Relações sintáticas	x	x	x	x	x		x			x	7
Estruturas sintagmáticas	x	x	x	x					x	x	6
Estruturas paradigmáticas		x	x	x				x	x	x	6
Thesauri	x	x	x	x	x					x	6
Métodos interpretativos	x	x			x			x	x		5
Classificação		x	x		x			x		x	5
Unitermo	x	x	x		x				x		5
Redes semânticas	x	x		x			x		x		5
Linguagem de informação		x	x	x	x		x				5
Instrumentos gramaticais		x	x		x		x	x			5
Indexação	x			x		x			x	x	5
Semântica					x		x	x	x	x	5
Indicador de função				x		x	x	x	x		5
Componentes de metalinguagem		x	x	x	x	x					5
Modelo de análise de documento					x		x	x	x	x	5

4.4 Discussão dos resultados

O número total de termos de indexação foi quatrocentos e oitenta e um (481), bastante elevado se considerarmos que correspondem a apenas um documento, embora seja relativamente longo pois tem trinta e duas (32) páginas. A indexação foi exaustiva, uma vez que não foram estabelecidos critérios mínimos nem máximos para a quantidade de termos. Os indexadores tiveram total liberdade na escolha dos termos, em um espaço de tempo considerável.

A baixa consistência foi ocasionada pela exaustividade, a ausência de qualquer restrição na indexação e a não utilização de um manual de indexação ou de vocabulários controlados. Também contribuíram alguns problemas de tradução, pois o artigo é em língua inglesa e alguns indexadores traduziram os mesmos termos diferentemente.

A falta de um manual que estabelecesse regras de indexação causou diferenças, sobretudo sintáticas como, por exemplo, cálculo de relação, cálculo relacionado, análise de documento, análisse

documentária, instrumentos semânticos e instrumentos de semântica. O poder de síntese e a linha de pensamento ou raciocínio difere de indexador para indexador, ocorrendo estruturas sintáticas diferentes: descritores, lista de descritores, SYNTOL, modelo SYNTOL, instrumentos semânticos, instrumentos semânticos de análise de conteúdo, tabulação, métodos de tabulação, modelos de procedimentos de análise de documentos, modelos de análise de documentos, ordem de entrada de índices, entradas de índices, índices, índice KWIC e assim por diante.

Na parte semântica houve maior concordância entre os indexadores, embora algumas palavras usadas isoladamente não apresentem significado para a indexação, como seleção, ordenação, ordem, ligação e outras. O problema inverso também ocorreu, isto é, termos compostos por mais de três palavras, formando verdadeiras frases. A média de consistência mais baixa, a do indexador F, foi causada exatamente por esse problema.

A média de consistência mais alta foi a do indexador H, cujo número de termos foi o menor do grupo, vinte (20).

A consistência de um indexador em relação aos demais (A → B, A → C, A → D e assim por diante) apresenta resultados diferentes da medida de cada indexador do grupo quanto a um único indexador (B → A, C → A, D → A e assim sucessivamente). Essa variação se deve aos diferentes números de termos indexados por cada indexador. Essa relação foi igual apenas para o indexador E (0,36).

A consistência recíproca (A ↔ D) foi igual apenas para os indexadores que selecionaram a mesma quantidade de termos: A e D (62 termos, consistência 0,33), B e E (48 termos e consistência 0,41), C e G (55 termos e consistência 0,21). No caso dos indexadores D e J, por apresentarem uma diferença de apenas um termo (62 termos para D e 63 para J), o resultado foi igual, embora na realidade haja uma diferença mínima (0,1774 e 0,1746).

Comparando-se a freqüência de termos e o título do artigo, Análise do Documento e Teoria Lingüística, verifica-se que análise de documentos apresentou freqüência 9 e teoria lingüística 7, portanto, freqüências bastante significativas. O termo de freqüência máxima foi metalinguagem (f.10).

A apresentação do artigo também fornece termos relevantes para indexação pois contém ciência da informação, lingüística, análise de documento, todos três, termos de freqüência 9.

5 CONCLUSÕES

Apesar da amostragem ser pequena quanto ao número de indexadores, apenas (10), e um único documento indexado, o número de termos analisados e medidos foi expressivo pois chegou a quatrocentos e oitenta e um (481).

Há uma tendência acentuada na escolha de termos únicos formados por apenas uma palavra e de termos compostos por duas ou três palavras no máximo, certamente pelo sentido natural de síntese da indexação.

A ausência de instrumentos de ajuda à indexação, como um manual que fornecesse regras ou vocabulário controlados contribuiu para tornar mais baixa a média de consistência.

Quanto maior a liberdade de indexação, maiores os riscos de inconsistência, porque a indexação exaustiva emprega grande número de termos. Conseqüentemente, a média de consistência mais significativa refere-se ao indexador que selecionou o menor número de termos (indexador H, 20 termos).

A indexação, embora seja um processo extremamente subjetivo, apresenta alta consistência em relação a termos que melhor caracterizam o conteúdo do documento. Um fator que deve ter contribuído para uma maior consistência é que dos dez indexadores, oito (8) pertencem à mesma área e todos são mestrandos de Ciência da Informação. Presume-se que por ser a grande maioria constituída de bibliotecários, que os indexadores possuam experiência em indexação e mesmo os estranhos à área, tenham algum conhecimento do assunto, ao menos teoricamente.

O resultado da freqüência indica que pode haver uma consistência absoluta de termos, o que é bastante improvável, até impossível, entre indexadores.

O indexador necessita de no máximo vinte minutos para indexação extraída de um texto relativamente longo, de trinta e duas (32) páginas, através da leitura do texto completo. Convém ressaltar que a indexação assim realizada, é muito dispendiosa para os sistemas de informção que, por esse motivo, geralmente não a adotam.

O estudo analisaria melhor a consistência e a economia da indexação se tivesse sido medida a consistência separadamente para cada tempo. Os termos, entretanto, não correspondem às mesmas partes do texto pois a leitura e a indexação variaram de indexador para indexador, dependendo da maior ou menor rapidez na realização da leitura e indexação. A cada dez minutos os indexadores faziam uma interrupção, partindo desse ponto para a indexação dos termos nos próximos dez minutos, daí a não coincidência dos tópicos indexados.

6 REFERÊNCIAS BIBLIOGRÁFICAS

DOYLE, L.B. *Information retrieval and processing*. Los Angeles, Melville Publishing C., 1975.

FOSKETT, A.C. *A abordagem temática da informação*. São Paulo, Ed. Polígono, 1973.

JONES, K.S. Index term weighting. *Information storage and retrieval*, 9:619-33, 1973.

JONES, P.E. & CURTICE, R.M. A framework for comparing term association measures. *American Documentation*, 13 (3), July, 1967.

MARON, M.E. & KUNHS, J.L. On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery*, 7 (3): 216-244, July 1960.

MARSHAK, J. Economics of inquiring, communicating, deciding. *American Economic Review*, 58: 697-706, 1968.

ROTHMAN, J. Index, indexer, indexing. In: KENT, A., LANCOUR, H., DAILEY, J.E. eds.— Encyclopedia of Library and Information Science, New York, 1974. v.11 p. 286-289.

SCHNEIDER, J.H. Selective dissemination and indexing of scientific information.

TINKER, J.F. Imprecision in meaning measured by consistency of indexing. *American Documentation*, 17 (2): 96-102, April 1966.

——— Imprecision in indexing. Part II. *American Documentation*, 19 (3): 322-30, July 1968.

ZUNDE, P. & DEXTER, M. Indexing consistency and quality. *American Documentation*, 20 (3): 289-67, July 1969.