Indexação: teoria e métodos

Giovana Deliberali Maimone
Nair Yumiko Kobashi
Denysson Mota

1. Introdução

A disciplina intitulada "Indexação: teoria e métodos" integra a grade curricular do curso de "Biblioteconomia e Documentação", oferecido pela Escola de Comunicações e Artes da Universidade de São Paulo. Tem como objetivo capacitar o aluno a compreender e desenvolver os processos de organização da informação, em particular a indexação, para recuperar informação em ambientes locais e eletrônicos. Além de extenso referencial teórico da área de Análise documentária, que engloba desde a leitura documentária até as normas, políticas e metodologias de indexação, a disciplina também conta com referências ao uso de softwares de indexação automática, de modo a promover a incorporação do ambiente eletrônico aos estudos do campo da Biblioteconomia. Visando a constante atualização do profissional da informação, são inseridas discussões sobre as tendências de indexação na web, como é o caso dos marcadores (tagging), das folksonomias e a otimização das tecnologias da informação para acesso e recuperação da informação. A prática de exercícios de indexação de materiais gráficos, em sala de aula, prioritariamente de textos científicos, possibilitam ao corpo discente desenvolver expertise técnica e intelectual no tratamento de documentos para indexação. Observa-se, neste sentido, a fundamental contribuição do estágio obrigatório

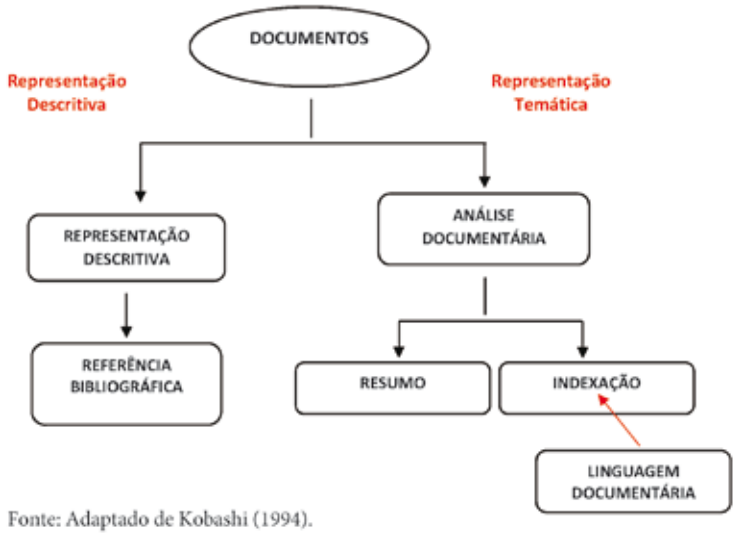
em Ambientes de Informação, verificando-se a aplicação prática das teorias e métodos propostos. Portanto, as atividades realizadas em sala de aula causam impacto significativo na elaboração de produtos documentários, função básica e essencial para a atuação competente do futuro profissional da informação.

2. Representação Documentária e Indexação

A representação de documentos é um conjunto de atividades que tem como principal intuito representar conteúdos informacionais de documentos, por meio dos elementos que os identifiquem de forma específica, para torná-los recuperáveis.

Deste modo, as ações de representação de documentos podem ser segmentadas em dois grandes grupos: a representação descritiva, que visa identificar um documento, a partir de suas características "físicas", e a representação temática, que supõe a análise, a condensação e a representação de documentos de acordo com os "assuntos" neles contidos. Os índices e os resumos documentários, são os principais produtos destas ações. Esses procedimentos podem ser representados graficamente como na figura abaixo:

Figura 1: *A representação de documentos*



Fonte: Adaptado de Kobashi (1994).

Ao tratar especificamente da indexação, pode-se caracterizá-la como um conjunto de ações que visam identificar e descrever conteúdos de documentos, de acordo com seus assuntos, por meio da "atribuição" de termos de uma linguagem de indexação, também chamada de linguagem documentária. De acordo com a norma "NBR 12676/1992 – Métodos para análise de documentos – Determinação de seus assuntos e seleção de termos de indexação", os estágios da indexação podem ser divididos em três: a) exame do documento e estabelecimento do assunto de seu conteúdo; b) identificação dos conceitos presentes no assunto; e, c) tradução desses conceitos nos termos de uma linguagem de indexação.

É essencial ressaltar que a indexação somente faz sentido se a recuperação da informação for efetiva, concorrendo, para isto, fatores que garantem sua qualidade, como por exemplo, a consistência na especificidade dos termos atribuí-

dos, o conhecimento e capacidade do indexador e a qualidade dos instrumentos de indexação utilizados (NRB 12676/1992).

Uma vez concebida a ideia de que a indexação é realizada por meio de atividades de análise, síntese e representação, pode-se adentrar nos aspectos que envolvem a leitura documentária, uma das primeiras variáveis que influem diretamente na qualidade da indexação acima comentada.

A leitura influencia e é influenciada pelos aspectos culturais e ideológicos que fazem parte da vivência humana, sendo que a interpretação de um texto depende do processamento das informações, do próprio texto de partida e do contexto de ambos (FUJITA, NARDI, SANTOS, 1998; FUJITA, 2004). Em leitura documentária, utiliza-se um procedimento específico de leitura chamada de "leitura técnica" que, segundo Cunha e Cavalcanti (2008, p. 222), compreende:

> [...] o estudo metódico do conteúdo de um documento que realiza o classificador para determinar os assuntos tratados e, mediante uma operação analítico-sintética, estabelecer os símbolos que representarão esse item no acervo. A leitura técnica compreende a análise do corpo central da obra, complementada por outras fontes de informação que integram a obra sob estudo, como o título e seu grau de representatividade, o sumário e o índice de assuntos, as orelhas e as contracapas, o prefácio, a catalogação e a classificação na fonte.

A leitura documentária, um dos aspectos da Análise documentária, tem como função principal condensar o conteúdo dos documentos para, em seguida, serem atribuídos a estes últimos os termos que expressam seu "assunto" ou tema. A leitura documentária depende, por sua vez, da qualidade do texto e da capacitação em realizar a análise, a síntese e a representação de textos. Neste contexto, é importante considerar algumas das condições impostas ao indexador. São elas: limite de tempo, propósito definido, geração de pro-

dutos, metodologia de indexação, contexto de trabalho, entre outras, sendo que uma das dificuldades apontadas pelos profissionais é compatibilizar o conteúdo informacional dos textos com os termos propostos pela linguagem do sistema (CINTRA, 1987, FUJITA, 1999).

A linguagem do sistema, que também podem ser chamadas de linguagem de indexação ou linguagem documentária, pode ser definida como: todo sistema artificial de signos normalizados, que facilitam a representação formalizada do conteúdo dos documentos para permitir a recuperação, manual ou automática da informação solicitada pelos usuários, sendo que seu objetivo principal é assegurar o controle de vocabulário de domínios de conhecimentos (CINTRA et al, 1997; FUJITA; LEIVA, 2010).

A função principal de uma linguagem de indexação é a de compatibilizar a linguagem utilizada por uma comunidade de usuários e entre várias instituições. Semelhante compatibilização é fundamental para elaborar estratégias de busca adequadas para a recuperação de informações (CINTRA et al, 1997). As linguagens organizadas como instrumentos de indexação e recuperação, de acordo com suas características, podem receber diferentes denominações: tesauros, vocabulários controlados, listas de cabeçalhos de assunto, taxonomias, ontologias, entre outras.

3. Política e Prática da Indexação

O principal propósito de um serviço de indexação é assegurar que documentos e informações cheguem aos usuários com precisão. Toda política de indexação é contextual, variando de acordo com o perfil da instituição e perfil e necessidades dos usuários.

Com o fim de tornar realizável a recuperação de informações presentes nos acervos dos diversos tipos de Centros Informacionais, uma política de indexação deve servir como

um guia para a tomada de decisões e deve levar em conta os seguintes fatores:

- Características e objetivos da instituição;
- Identificação dos usuários;
- Identificação de recursos humanos, materiais e financeiros que delimitam o funcionamento de um sistema de recuperação de informações.

Verificadas as necessidades particulares de cada instituição, é possível realizar indexações mais condizentes com as necessidades dos respectivos públicos. Neste sentido, é pertinente avaliar as políticas de indexação, sendo que alguns elementos devem ser considerados, como:

- Aspecto estratégico: para quem a política é direcionada? Quais os limites de sua aplicação?
- Ferramentas de indexação: quais são as linguagens e esquemas de classificação utilizados? Quais são as características de cada linguagem?
- Aplicação das ferramentas: como serão utilizadas? Quais são os tipos de documentos indexados?

Em complemento ao exposto acima, pode-se dizer que na avaliação de um Sistema de Indexação e Resumos (SIR) procura-se determinar a qualidade do produto oferecido, o seu nível de desempenho em relação às necessidades dos usuários da informação e os custos decorrentes. Assim, é fundamental, segundo Lopes (1985), analisar os seguintes aspectos:

- Relevância = capacidade do sistema em fornecer respostas que realmente correspondam à questão proposta.
- Cobertura = abrangência em relação à literatura sobre um assunto. Indexar todos os materiais publicados é impossí-

vel e a cobertura é determinada pela proporção da literatura sobre um tema, ou temas, incluída no SIR.

- Revocação = capacidade do SIR de oferecer, em resposta a uma questão, todas as referências relevantes existentes na base de dados. Esta é uma questão controvertida. Muitos teóricos da área consideram difícil avaliar a revocação dos sistemas.
- Precisão = capacidade do SIR em fornecer apenas referências relevantes, eliminando as que não são relevantes para a questão proposta.
- Novidade = medida da proporção de referências relevantes recuperadas no SIR, que o usuário não conhecia anteriormente.
- Esforço do usuário = dependente da precisão com que os materiais são indexados.
- Tempo de resposta = tempo decorrido entre a pergunta e a resposta fornecida pelo sistema.
- Produtos oferecidos = são compostos por relatórios de bases de dados, índices impressos, resumos e as formas de saída que resultam das buscas de informação (qualidade tipográfica, existência de resumos, referências, etc.).
- Linguagem de indexação = usada para identificar os assuntos dos documentos, sendo que a qualidade da indexação tem reflexos diretos na recuperação de informações e satisfação dos usuários.

É necessário enfatizar que a escolha da linguagem documentária é fator essencial para a eficácia de um sistema de recuperação da informação, que deve considerar, principalmente: os objetivos do sistema, o tipo de usuário e a abrangência / especificidade dos assuntos dos documentos armazenados no sistema. Neste sentido, podem ser destacadas: as linguagens de indexação pré-coordenadas, que combinam ou coordenam os termos no momento da indexação e as pós-

coordenadas, que combinam ou coordenam os termos no momento da busca (VALE, 1987).

De modo geral, um índice documentário é um mecanismo ou instrumento auxiliar, usado, tanto na armazenagem, como na busca e recuperação da informação. O índice é utilizado, via de regra, para localizar um material numa base de dados, como instrumento de indexação para auxiliar o indexador no momento da "escolha" dos termos a serem atribuídos a um documento, como ferramenta para auxiliar o usuário a encontrar o material de que necessita, recorrendo aos descritores das linguagens documentárias, como também, para a aprendizagem de novos termos e relações entre eles (CUNHA, CAVALCANTI, 2008).

Além da indexação, dita "manual", também há outros dois tipos de indexação, chamadas de "automática" e "semiautomática".

4. Indexação Automática e Semiautomática

A indexação não ocorre apenas de forma manual. Embora denominada comumente de indexação "manual", é melhor caracterizá-la como uma operação intelectual humana. É possível realizar a indexação também de forma assistida por computador. Lancaster (2004) aborda as diversas formas de tratamento para a geração de termos de indexação e resumos, seja de forma automática ou semiautomática. Para o autor, as diferentes formas e técnicas de indexação utilizadas auxiliam o profissional da informação em seu trabalho diário, automatizando um trabalho, de certa forma, repetitivo.

A forma mais simples de indexação automática consiste em contar palavras, um mecanismo que simula a leitura humana pelo computador. As contagens são geralmente realizadas mediante o uso de duas listas: uma negativa e uma positiva. A primeira contém palavras consideradas pouco significativas que, por isso, precisam ser ignoradas, como os

pronomes ou outras palavras consideradas vazias de significado para os fins da indexação. A segunda lista é constituída de palavras consideradas muito relevantes e que, caso apareçam no texto analisado, serão contabilizadas. O software SISA (GIL LEIVA, 2003) opera, como muitos softwares, com essas duas listas para gerar as recomendações de palavras-chave.

É possível também que estes tipos de softwares contenham palavras que não estão em nenhuma das duas listas, isto é, são listagens de palavras que podem ser relevantes mas que não foram consideradas previamente, principalmente na lista positiva. Além de usar palavras, também é possível extrair sentenças completas que sejam julgadas relevantes para identificar e descrever o texto analisado (LANCASTER, 2004). Estes mecanismos, no entanto, exigem maior trabalho nas etapas de elaboração do projeto e programação, principalmente para estabelecer a frequência de aparecimento das palavras que possam ser consideradas relevantes nos documentos, assim como para definir uma lista das expressões que precisam ser avaliadas.

O principal aspecto da indexação automática é o cálculo ou estabelecimento da relevância das palavras e expressões. A frequência da palavra no texto, de forma pura e simples, pode ser, em diversos casos, enganadora, levando o software a considerar como relevante uma palavra pouco significativa. Lancaster (2003) usa o exemplo da palavra *biblioteca* em um acervo sobre Biblioteconomia para ilustrar que uma palavra muito frequente, em uma base ou repositório, em que ela aparece bastante, torna-se irrelevante. Nesse sentido, a relação entre a frequência das palavras em um texto e sua frequência no acervo, portanto a frequência relativa de um termo, é considerada mais rica e eficaz em termos de indexação automática. Para ilustrar, Lancaster (2004) utiliza

a palavra *amianto* em um acervo de Biblioteconomia, ou a palavra *biblioteca* em um acervo de uma fábrica de cimento.

O software, depois de realizada a contagem, o sistema apresenta ao indexador (no caso da indexação semiautomática) ou vincula diretamente (no caso da indexação totalmente automática) os termos de indexação selecionados como relevantes. Essa vinculação pode ocorrer de forma direta, isto é, as palavras vinculadas são apenas as extraídas do texto, ou indireta, por atribuição. Este último modo se assemelha à forma com que os profissionais de biblioteconomia trabalham, escolhendo termos considerados como descritores do texto, mesmo que elas não estejam materialmente expressas no texto. O computador realiza essa atribuição mediante a comparação das palavras contabilizadas com uma lista de termos relacionados ou equivalentes, atribuindo não apenas os termos encontrados, mas aqueles que aparecem na lista. O software SISA também possibilita incluir uma lista de termos gerais para uso nestes moldes (GIL LEIVA, 2003). Lancaster (2004) usa o exemplo do termo *chuva ácida*, que poderá ocorrer, por exemplo, na atribuição de termos relacionados a *poluição atmosférica* e *dióxido de enxofre*.

No entanto há problemas com a indexação automática. Os mais comuns são a subatribuição, situação em que o computador não seleciona os termos que os indexadores humanos atribuiriam ao texto, e a superatribuição, quando o computador seleciona termos que os indexadores humanos não considerariam relevantes para representar o texto. Segundo Lancaster (2004), os termos escolhidos por computadores são de 80% a 90% semelhantes aos selecionados por humanos. Portanto, há uma faixa de 10% a 20% de termos que ou foram atribuídos desnecessariamente, ou não foram identificados como relevantes.

5. Considerações Finais

A indexação, como exposto neste texto, é fundamental para que ocorra a adequada recuperação de informações. Certamente, é necessário buscar formas cada vez mais aprimoradas de realizar esse processo, seja pela intervenção direta de indexadores humanos, seja por meio de procedimentos assistidos por computador. Procurou-se, aqui, expor os principais aspectos da indexação, tais como a análise, a condensação e a representação de conteúdos informacionais. A análise é um processo de interpretação que, na indexação, requer técnicas específicas apoiadas nas teorias de processamento de textos e na Linguística textual (CINTRA, 1987). A condensação, por sua vez, é um procedimento intelectual necessário à compreensão de textos. Teun Van Djik e Walter Kintsch (1983) são os autores que sistematizaram o processo de compreensão de textos, afirmando que nessa operação cognitiva ocorrem: a) o apagamento de informações consideradas pouco relevantes, b) a seleção de informações relevantes; c) a condensação por generalização e a reescritura condensada do conteúdo pertinente identificado. No caso da indexação, a tradução do enunciado reescrito, em descritores padronizados, é a operação final.

A indexação automática combina conhecimentos linguísticos e computacionais para a elaboração de sistemas de processamento de textos. É um procedimento muito útil que necessita, no entanto, de aprimoramento para ser efetivamente utilizado nos sistemas de informação.

Deve-se assinalar, por fim, a importância das linguagens documentárias para representar os conteúdos dos documentos. Essas linguagens são linguagens artificiais, construídas especificamente para indexar e buscar informação em sistemas de informação. Há diferentes os tipos de linguagens utilizadas nos Sistemas de informação, tais como os sistemas de classificação, os tesauros, as taxonomias, as ontologias.

Cada tipo de linguagem serve a uma finalidade. Portanto, saber escolher a linguagem adequada para o sistema de informação é uma das tarefas importantes do indexador. As linguagens documentárias são, por outro lado, ferramentas dinâmicas que devem ser atualizadas constantemente, em sintonia com a criação de novos conceitos pelos diferentes domínios do saber. É importante, portanto, que o indexador conheça os procedimentos de atualização de linguagens para que essas ferramentas mantenham sempre as suas funcionalidades como instrumentos de controle de vocabulário.

Muitos são os conhecimentos necessários para que a indexação se realize de forma adequada. Cabe, portanto, aos profissionais da informação estarem atentos para as mudanças nas teorias, nos métodos, nas ferramentas e nos recursos computacionais criados pelas áreas que se dedicam à organização da informação e do conhecimento.

Referências

CARNEIRO, M. V. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 221 – 241, set. 1985.

CINTRA, A. M. M. Estratégias de leitura em documentação. In: SMIT, J. W. (coord.) **Análise documentária***:* a análise da síntese. 2. ed. Brasília, DF: IBICT, 1987.

CINTRA, A. M. M. et al. **Para entender as linguagens documentárias**. 2. ed. rev. e ampl. São Paulo: Polis, 2002. 92 p.

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de biblioteconomia e arquivologia**. Brasília, DF: Briquet de Lemos, 2008.

FUJITA, M. S. L. A leitura do indexador: estudo de observação. **Perspectivas em Ciência da Informação***,* Belo Horizonte, v. 4, n. 1, p. 101 – 116, jan. / jun. 1999.

FUJITA, M. S. L. A leitura documentária na perspectiva de suas variáveis: leitor-texto-contexto. **Datagramazero – Revista de Ciência da Informação**, v. 5, n. 4, ago. 2004.

FUJITA, M. S. L.; LEIVA, I. G. As linguagens de indexação em bibliotecas nacionais, arquivos nacionais e sistemas de informa-

ção na América Latina. XVI Seminário Nacional de Bibliotecas Universitárias. **Anais**. 2010.

FUJITA, M. S. L.; NARDI, M. I. A.; SANTOS, S. A leitura em análise documentária. **Transinformação**, v. 10, n. 3, p. 13 – 31, set./dez. 1998.

GIL LEIVA, I. Sistema para la Indización SemiAutomática (SISA) de Artículos de Revista de Biblioteconomía y Documentación. In: II Jornadas de Tratamiento y Recuperación de Información, 2003, Leganés (Madrid), **Anais eletrônicos**… Leganés (Madrid), 2003. p. 228-232. Disponível em: http://webs.um.es/isgil/SISA IndizacionautomaticaAutomaticIndexingGI LEIVA.pdf. Acesso em: 23 nov. 2004.

KOBASHI, Nair Yumiko. **A elaboração de informações documentárias***:* em busca de uma metodologia. São Paulo: ECA/USP, 1994. 195 p. (Tese de doutorado).

LANCASTER, F. W. **Indexação***:* teoria e prática. 2. ed. rev. atual. Brasília, DF: Briquet de Lemos/Livros, 2004.

LOPES, E. de F. Avaliação de serviços de indexação e resumo: critérios, medidas e metodologia. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 242-255, set. 1985.

LOUSADA, M. *et al.* Políticas de indexação no âmbito da gestão do conhecimento organizacional. **Informação e Sociedade***:* estudos. João Pessoa, v. 21, n. 1, p. 191 – 202, jan./abr. 2011.

NBR 12676/1992. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: Métodos para análise de documentos – Determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992.

RUBI, M. P.; FUJITA, M. S. L. Elementos de política de indexação em manuais de indexação de sistemas de informação especializados. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 8, n. 1, p. 66 – 77, jan./jun. 2003.

SMIT, J. W. **Análise documentária***:* a análise da síntese. Brasília: IBICT, 1987. 133 p.

VALE, E. A. do. Linguagens de Indexação. In: SMIT, Johanna W. (org.). **Análise documentária***:* a análise da síntese. Brasília: Ibict, 1987. 133 p.

VAN DJIK, T.; KINTSCH, W. **Strategies of discourse comprehension**. Orlando: Academic Press, 1983.