INSTRUMENTOS E METODOLOGIAS DE REPRESENTAÇÃO DA INFORMAÇÃO

MARIA SALET FERREIRA NOVELLINO

Resumo
Contextualiza e conceitua o processo de representação da informação. Analisa a evolução histórica dos instrumentos que foram e vêm sendo utilizados para a representação da informação. Apresenta uma descrição das várias concepções que subjazem às metodologias para a representação da informação.

Palavras-Chave
Análise de assunto. Indexação de assunto. Linguagem documentária. Representação da informação. Tesauro. Transferência de informações.

1 Introdução

A Ciência da Informação é uma disciplina voltada para o estudo de fenômenos subjacentes à produção, circulação e uso da informação. O estudo desses fenômenos tem como finalidade possibilitar a criação de instrumentos e o estabelecimento de metodologias que viabilizem a transferência de informações.

O conceito de transferência de informações é aqui compreendido como a intervenção realizada por sistemas de organização do conhecimento e recuperação da informação em determinadas ações comunicativas, que são aquelas que se dão entre produtores e consumidores de conhecimento.

As ações comunicativas, as quais têm como instrumento a linguagem, podem se realizar mediante a relação (a)entre falante e ouvinte; (b)entre imagem e aquele que assiste (c)entre texto e leitor. A Ciência da Informação volta-se, principalmente, para a ação comunicativa entre texto e leitor, tendo como objetivo principal criar condições para a sua realização. Ela intervém na ação comunicativa textual para garantir que ela efetivamente ocorra, isto é, que a informação torne-se acessível àquele que dela precisa.

A intervenção na ação comunicativa pode ser abordada de duas maneiras distintas: (a) sob o ponto de vista da recuperação da informação; ou (b) sob o ponto de vista da representação da informação.

Sob o ponto de vista da recuperação da informação, ênfase é dada à abordagem cognitiva, na qual a compreensão que o usuário tem de

determinadas disciplinas ou áreas de assunto prevalecem, bem como seu comportamento no que diz respeito à busca por informações. As pesquisas desenvolvidas nessa área voltam-se para a saída do sistema, preocupando-se em estabelecer interfaces amigáveis e permitindo que o usuário interfira na linguagem do sistema (user-modelling). Vêm desenvolvendo-se homologamente à Ciência da Computação, aplicando princípios e métodos da inteligência artificial e da lingüística computacional.

Sob o ponto de vista da representação da informação, ênfase é dada à organização do conhecimento. A organização do conhecimento no âmbito da Ciência da Informação diz respeito ao desenvolvimento e avaliação de teorias para análise de determinadas áreas de assunto visando a elaboração de instrumentos e métodos para a representação das informações geradas nessas áreas de assunto. As pesquisas desenvolvidas nessa área voltam-se para a entrada do sistema. Desenvolve-se homologamente à compreensão científica de estrutura do conhecimento, projetando metodologias para a análise de assunto e para a geração de sistemas de classificação e linguagens documentárias

O nosso objeto de estudo é a transferência de informações desde um ponto de vista da representação da informação.

2 Representação da Informação

A principal característica do processo de representação da informação é a substituição de uma entidade lingüística longa e complexa - o texto do documento - por sua descrição abreviada. O uso de tal sumarização não é apenas uma conseqüência de restrições práticas quanto ao volume de material a ser armazenado e recuperado. Essa sumarização é desejável pois sua função é demonstrar a essência do documento. Ela funciona então como um artifício para enfatizar o que é essencial no documento considerando sua recuperação, sendo a solução ideal para organização e uso da informação.

O processo de representação da informação envolve dois passos principais:

1) análise de assunto de um documento e a colocação do resultado desta análise numa expressão lingüística.

2) atribuição de conceitos ao documento analisado.

A realização desta última fase pressupõe uma linguagem documentária, instrumento de padronização da indexação, a qual visa garantir que indexadores de um mesmo sistema ou sistemas afins usem os mesmos conceitos para representar documentos semelhantes. Ela também é um instrumento de comunicação ao permitir que indexadores e usuários partilhem um mesmo vocabulário.

3 Instrumentos para a Representação da Informação

Nem sempre as linguagens documentárias foram pensadas como instrumentos de indexação e recuperação. Inicialmente, elas tinham como objetivo apenas padronizar as entradas de assunto de catálogos ou índices. As primeiras foram as listas de cabeçalhos de assunto. Houve, a seguir, a adesão ao vocabulário livre, isto é, a opção pela ausência

de um controle do vocabulário usado para a indexação. Voltou-se, posteriormente, ao controle do vocabulário, empregando-se as listas de termos autorizados. A preocupação com a criação de um instrumento de representação da informação voltado para a recuperação, e, conseqüentemente, para demonstrar ao usuário a estrutura da linguagem de representação deu origem aos tesauros, tesauros facetados e classauros.

As listas de cabeçalhos de assunto foram construídas para instrumentalizar a indexação de assuntos de documentos, que seriam registradas em fichas catalográficas para compor o catálogo alfabético de assuntos. Elas foram projetadas para bibliotecas de acervos gerais e compreendiam o conhecimento como um universo fragmentável em disciplinas.

As críticas às listas de cabeçalhos de assunto e aos sistemas pré-coordenados, nos quais eram utilizadas, eram as seguintes:

a) Impossibilidade de acesso direto aos subcabeçalhos, o que significava a inacessibilidade a uma série de conceitos.

b) As listas enumeravam conceitos que deveriam ser usados tal como nelas apareciam e a inserção de novos termos dependia de uma garantia literária, o que comprometia a especificidade da linguagem.

c) A representação verbal e notacional exigiam uma demanda de tempo tal, que sempre haveria um volume considerável de material por ser processado e, portanto, irrecuperável.

A necessidade de tratar tematicamente a informação de uma forma mais específica devida à especialização dos acervos, e de criar formas de representação/recuperação mais ágeis, devida ao tipo de material armazenado, relegou a um segundo plano, e, em casos mais radicais, levou ao abandono do controle do vocabulário.

A opção pelo vocabulário livre foi característica dos primeiros sistemas pós-coordenados, que não reuniam os termos no momento da indexação nem estabeleciam assunto principal. Atribuíam-se quantos termos isolados fossem necessários para descrever determinado documento cabendo ao usuário coordená-los no momento da busca.

O primeiro sistema pós-coordenado, o Unitermo, foi previsto para uso em fichas. Com a introdução do computador nos sistemas de recuperação da informação, os sistemas pós-coordenados começaram a ser massivamente empregados em sistemas especializados.

O computador foi introduzido na área inicialmente para a produção de índices impressos: ordenação automática dos termos e títulos. Mas com o desenvolvimento tecnológico, o computador tornou-se instrumento não só para a produção e a compilação de índices, mas também para a geração dos próprios índices: extração e atribuição de palavras ou conceitos. Houve como conseqüência uma implementação de sistemas de indexação pós-coordenados. O computador permitia o uso da lógica booleana, lógica de combinação binária por soma, produto ou diferença que se ajustava à coordenação de termos no momento da recuperação.

Essa vantagem da rapidez no tratamento da informação começou a ser derrubada pelas desvantagens apresentadas na recuperação de

documentos: multiplicidade de termos para representar um mesmo conceito, descontextualização dos termos em relação ao assunto total do documento e também da área de domínio da qual fazia parte. Este problema foi identificado como causado pela ausência de um instrumento que padronizasse as linguagens dos produtores e usuários da informação.

Voltou-se ao controle do vocabulário. Para evitar falsas coordenações passou-se a usar conceitos pré-coordenados e para evitar o uso de várias palavras para um mesmo conceito, relacionamentos de equivalência passaram a ser estabelecidos. Surgem então as listas de termos autorizados, que continham registros de decisões tomadas, em relação à indexação, como modelos para os indexadores. Isto é, era um registro de tomada de decisões no que diz respeito à seleção de conceitos para indexação.

Quando, porém, estes instrumentos começaram a ser projetados não mais apenas como auxiliares da indexação mas também como da recuperação, outros mecanismos associativos passaram a ser considerados. Surgiram então os tesauros, que adicionaram a este relacionamento entre os termos de indexação, outros, visando instrumentalizar não só a representação mas também a busca da informação.

A Lista, desde a sua gênese, foi pensada apenas enquanto instrumento padronizador da indexação. Já o Tesauro, foi idealizado como instrumento facilitador da comunicação dentro do sistema, padronizando as linguagens de indexação e de recuperação, a partir da terminologia da área representada. Mas a ordenação puramente verbal (alfabética) dos tesauros levaram a que se perdesse ou se deixasse de demonstrar, para os usuários do instrumento, a estrutura classificatória nele embutido. Para o estabelecimento de relações genéricas (hierárquicas e partitivas) e associativas quando da construção dos tesauros, características de divisões são estabelecidas. Na ordenação alfabética final, as categorias de conceitos que nortearam a divisão e os termos a elas subordinados perdem-se, pois não ficam explícitas no texto final do tesauro. A necessidade de deixar explícita a organização de determinadas áreas de assunto conduziu ao tesauro facetado e ao classauro.

O tesauro facetado e o classauro apresentam duas ordenações: a alfabética e a classificada, o que permite tornar visível ao usuário do tesauro, seja ele o indexador ou o usuário do sistema, a classificação a ele subjacente e que antes só era clara aos elaboradores do instrumento. Eles surgiram como tentativa teórica e prática de reunir as vantagens da linguagem documentária verbal e dos sistemas de classificação facetados, assumindo que a teoria da classificação facetada seria a base para a estruturação de uma linguagem documentária verbal.

Fugmann (1), enumerou as vantagens dos sistemas de classificação e dos tesauros:

A grande vantagem de um sistema de classificação é que, nele, as características de divisão que nortearam o classificacionista ficam visíveis. Os meta-conceitos são apresentados no corpo da tabela, demonstrando ao usuário a forma de organização daquela área do conhecimento. Além disso, num esquema de classificação, os conceitos

que são subordinados a um conceito mais geral podem ser agrupados mais corretamente de acordo com a característica de divisão que guiou esta reunião. Características de divisão dão ao vocabulário transparência e assim enriquecem a busca, localizando e relacionando o conceito de acordo com suas características intrínsecas.

Uma das vantagens do tesauro é a possibilidade de expressar o conjunto completo de relações associativas entre conceitos e não apenas relações genéricas. Além disso, indexadores e usuários estão mais familiarizados com os termos expressos em linguagem natural de um tesauro do que com as notações de um sistema de classificação. Um sistema de classificação e um tesauro usados concomitantemente seriam complementares um ao outro. Lançando-se mão das vantagens que cada um oferece, controlar-se-ia os pontos fracos que cada um apresenta. Teríamos então os tesauros facetados e os classauros.

As linguagens documentárias verbais vêm se aproximando, cada vez mais, da teoria da classificação. Elas que, em sua origem, pareciam negar os princípios classificatórios, buscam hoje nesta teoria fundamentos para a organização de conceitos que transcendam as limitações do arranjo verbal. A necessidade de transparência da organização do vocabulário estimulou o aparecimento dos tesauros facetados e dos classauros.

No início da história do controle do vocabulário, linguagens verbais e notacionais eram independentes. A automatização da indexação, de início, resultou num privilegiamento da verbal porém sem controle ou padronização e numa subvalorização da notacional. Posteriormente, passa-se a buscar coordenar as linguagens verbais e notacionais num só instrumento: tesauros facetados e classauros. Além disso, a teoria da classificação é fortalecida como paradigma para a análise conceitual de áreas de assunto.

Fatores que contribuíram para a valorização da teoria da classificação:

a) Linguagem documentária como instrumento de busca e a conseqüente necessidade de apresentar ao usuário a estrutura/classificação daquela área do conhecimento de modo que pudesse desenvolver sua busca (a preocupação com a padronização da representação deixa de ser primordial).

b) O acesso direto à coleção não através da estante mas mediante uma tela que exponha não só os itens sob cada conceito recuperado mas que o contextualize no universo do conhecimento sob interesse de pesquisa (a preocupação com a notação para armazenamento deixa de ser primordial).

Com isso, a classificação como recurso para padronização e guarda/endereçamento de livros perde sua importância colocando a classificação, mais propriamente, a teoria da classificação, não como instrumento mas como base para análise, representação e busca da informação.

4 Metodologias para a Representação da Informação

A partir de determinadas compreensões do significado de "assunto", procedimentos para identificá-lo são estabelecidos. Esses procedimentos vêm a compor as metodologias para a representação da informação.

Lançando mão de modelo construído por Albrechtsen(2), apresentaremos, abaixo, concepções metodológicas para a representação da informação:

(a) concepção simplística: vê os assuntos como entidades absolutas objetivas que podem ser derivadas como abstrações lingüísticas diretas de documentos ou resumidas como cifras (figures) matemáticas, usando métodos de indexação estatística. De acordo com esta concepção, a indexação pode ser totalmente automatizada. A concepção simplística de análise de assunto vê os assuntos como abstrações diretas dos documentos. Seguindo esta concepção, extrair-se-ia automaticamente todas as palavras ou expressões dos textos.

(b) concepção orientada ao conteúdo: envolve uma interpretação dos conteúdos dos documentos que vão além do léxico e algumas vezes da estrutura superficial gramatical, que é o limite dentro do qual a concepção simplística opera. A análise de assunto dos conteúdos dos documentos envolve a identificação de tópicos ou assuntos que não são explicitamente colocados na estrutura textual superficial de um documento, mas que são prontamente perceptíveis por um indexador. Conseqüentemente, envolve uma abstração mais indireta do próprio documento. A concepção orientada ao conteúdo baseia-se tanto nas informações explícitas quanto nas implícitas presentes nos textos. Por informação de assunto explícita entende-se informação que é expressa na terminologia aplicada pelo produtor do documento. Um documento pode também trazer informação implícita, a qual não é diretamente expressa pelo autor, mas é prontamente compreendida ou interpretada pelo leitor (humano) de um documento. Esta é a abordagem mais comum para a indexação de assuntos. Entretanto, ela se limita a representar ou resumir o documento como uma entidade isolada. A análise de assunto focaliza o documento como uma fonte isolada de conhecimento, embora o indexador seguindo esta concepção possa considerar o contexto do documento: a coleção a qual ele pertence (intertextualidade).

(c) concepção orientada à necessidade: vê as entradas de assunto (subject data) como instrumentos para a transferência de conhecimento. Tendo como objetivo, conseqüentemente, localizar informação pragmática ou conhecimento. De acordo com esta concepção, os documentos são criados para a comunicação do conhecimento, e as entradas de assunto deveriam ser feitas para funcionar como instrumentos para mediar e traduzir este conhecimento visível para quaisquer pessoas interessadas. A concepção de análise de assunto orientada à necessidade aplica-se aqui como um denominador comum para abordagens orientadas à necessidade (request) e esquemas (frameworks) sociológico-epistemológicos para a indexação. A análise de assunto, baseada na necessidade, vincula um foco diferente da análise de assunto orientada ao conteúdo. Ao analisar um documento, o indexador não se concentra na representação ou resumo das informações explícitas e implícitas, mas pergunta: como posso tornar este documento ou parte dele visível aos usuários em potencial? Quais termos devo usar para levar este conhecimento àqueles interessados?

Na indexação orientada às necessidades são

as buscas dos usuários por conhecimento em sistemas de recuperação da informação ou índices que determinam o método de indexação. Portanto, um documento é analisado com o propósito de predizer sua potencialidade para atender a grupos particulares de usuários.

Hjorland (3) analisou as várias formas de tratar o conceito "assunto" em Ciência da Informação, e as caracterizou da seguinte maneira:

(a) Concepção ingênua (naive): para a qual estabelecer o assunto de um documento não constituiria problema, pois seria um processo óbvio: o título daria a indicação necessária.

(b) Idealismo subjetivo: toma conceitos e assuntos para expressarem as percepções ou visões de um ou mais indivíduos. Aqui, a chave para o conceito de assunto repousa no estudo das mentes de algumas pessoas: autores ou usuários de documentos.

(c) Idealismo objetivo: enquanto o idealismo subjetivo super-enfatiza as percepções, o idealismo objetivo tende a super-enfatizar certos aspectos de uma análise teórica e torná-los absolutos. As idéias existiriam fora da consciência humana, a priori, assim como são a priori dos conceitos expressos nos documentos. Estas idéias ou assuntos teriam propriedades universais ou fixas, podendo ser analisadas num sistema universal ou separado em partes individuais. O idealismo objetivo se expressa num processo de classificação com a visão de que a classificação de documentos poderia ser feita independentemente do contexto no qual a classificação está sendo usada.

(d) Conceito pragmático de assunto: nesta visão, os documentos são indexados para serem recuperados. A indexação não se orienta pelo conteúdo mas pela demanda. Uma indexação orientada à necessidade ou ao usuário, é a descrição de um assunto o qual pode ser percebido como a relação entre as propriedades de um documento e uma necessidade do usuário real ou antecipada.

(e) Teoria de assuntos realista/materialista: de acordo com esta abordagem, os documentos são um problema teórico. De uma lado, os documentos refletem a visão subjetiva do autor dos assuntos tratados, e de outro lado, o documento tem propriedades objetivas , que seriam toda proposição (statement) verdadeira que pode ser dita sobre o documento. Essas propriedades emergem especialmente no uso do documento . Por exemplo, lendo um documento em conexão com uma atividade em particular: pesquisa, educação, etc.

Hjorland e Albrechtsen identificam e enumeram as tendências em representação da informação e apresentam uma tendência emergente que se enquadra na concepção orientada à necessidade e na teoria de assuntos realista/ materialista: a análise de domínio. (4)

A análise de domínio é uma metodologia para construção de modelos de representação da informação a partir da investigação de determinadas características de domínios específicos do conhecimento: a identificação de condições culturais, históricas e lingüísticas que imponham exigências particulares para a construção de modelos de domínio tais como sistemas de classificação ou tesauros. Compreende também um ponto de vista epistemológico para identificar os paradigmas

científicos e técnicos, abordagens de pesquisa e interesses de conhecimento nos domínios cobertos. (5)

A análise de domínio, como os sistemas de classificação tradicionais, volta-se para o estudo e análise da estruturação de áreas de assunto, porém com uma diferença fundamental: vai lidar com a contextualização dos conceitos na sua área de domínio, na qual o documento não deve ser interpretado como uma fonte isolada de conhecimento, mas como parte de uma área de conhecimento, uma contribuição a ela. A análise de domínio procura contextualizar a representação mais amplamente, considerando não só a terminologia empregada em determinada área de assunto, ou os termos ocorrentes na literatura da área; mas também o uso que é feito da informação produzida, isto é, a sua aplicação para a elaboração de serviços e produtos; as pesquisas desenvolvidas, que representam os caminhos para onde a área de assunto analisada vai progredindo; o ensino, que significa o conhecimento já estabelecido na área.

Referências

(1) FUGMANN, R. An interactive classaurus on the PC.*International Classification*, v.17, n.3/4, p.133-137,1990.

(2) ALBRECHTSEN, Hanne. Subject analysis and indexing: from automated indexing to domain analysis. *The Indexer*, v.18, n.4, p.219-224, October 1993

(3) HJORLAND, Birger. The concept of 'subject' in Information Science. *Journal of Documentation*, v.48, n.2, p.172-200, June 1992

(4) _____, ALBRECHTSEN, Hanne. Toward a new horizon in information science: domain analysis. Journal of the American Society for Informacion Science, v.46, n.6, p.400-425, 1995.

(5) ALBRECHTSEN, Hanne (moderator). Domain analysis in Information Science: investigations into the nature and structure of knowledge domains for classification and retrieval. In: *Proceedings of the 56th ASIS Annual Meeting*. v.30, 1993. p.290-291.

Maria Salet Ferreira Novellino
Mestra em Ciência da Informação UFRJ/IBICT.
Doutoranda em Ciência da Informação UFRJ/IBICT.

Title
Tools and Methodologies for Information Representation

Abstract
In this article, the information representation process is put in context and conceptualized , and its instruments are historically analyzed. It also presents a description of several conceptions that underlie the methodologies for information representation.

Keyworlds
Subject analysis. Subject indexing. Documentary languages. Information representation. Thesaurus.Information transfer.

Artigo recebido em 05/06/96