A RELAÇÃO INESCUSÁVEL ENTRE LINGÜÍSTICA E DOCUMENTAÇÃO

Leilah Santiago Bufrem
Professora Titular do Departamento de Ciência e Gestão da Informação
Universidade Federal do Paraná
bufrem@milenio.com.br

MOREIRO GONZÁLEZ, José Antonio. **El contenido de los documentos textuales**: su análisis y representación mediante el lenguaje natural. Gijón: Ediciones Trea, 2004.

Se as transformações sofridas pelos modos de criar, registrar, reproduzir e acessar o conhecimento são determinadas historicamente, permanece imutável a expressão e posterior descrição dos conceitos mediante as palavras. Estas são apresentadas como referência inevitável nos processos e resultados da análise de conteúdo documental por Moreiro González, em sua mais recente obra, ***El contenido de los documentos textuales***: *su análisis y representación mediante el lenguaje natural.* Seja utilizada livremente, como na linguagem natural, seja controlada por um sistema que a legitima para representar conceitos, a palavra converte-se em fim e meio para o autor, o leitor e o analista. Essas e outras considerações preliminares sobre os desdobramentos fenomenológicos da palavra introduzem o leitor ao campo da análise e representação documentais do conteúdo dos textos.

Já conhecido internacionalmente no mundo da Biblioteconomia e Ciência da Informação, José Antonio Moreiro González é especialmente valorizado no Brasil, para onde tem vindo com freqüência, sempre a convite de instituições e colegiados do mundo acadêmico, a fim de ministrar cursos, realizar palestras e conferências e participar de eventos culturais. Em nossos meios universitários, é conhecida sua alentada produção intelectual, destacando-se o título **O conteúdo da imagem**, livro que tem sido adotado em cursos de graduação e pós-graduação no país, traduzido para o português e publicado pela Editora da Universidade Federal do Paraná.

Desta vez, ao tratar da análise e representação do conteúdo de documentos textuais, Moreiro parte da exigência da determinação das estruturas semânticas desses documentos para conhecer sua organização e discriminar as partes em que se concentra a informação relevante. Pautando-se na concepção de Báez, sobre a intencionalidade do ato da comunicação, dirige-se a um objeto semântico, a informação, expressa pela linguagem em articulações concretas, denominadas de documentos.

A análise de conteúdo e a intermediação documental são enfocadas no primeiro capítulo da obra e reconhecidas as barreiras à comunicação direta das mensagens. Com apoio no dualismo da teoria saussuriana, explicita-se a associação de interdependência compositiva

dos documentos, com suas duas estruturas, a externa e a abstrata, compreensíveis para o leitor documentarista que, sem atuar como autor da mensagem, torna-se emissor ou viabilizador da intermediação necessária entre ela e seu destino. Essa prática reconhece também níveis descritivos e de análise que introduzem o leitor ao que o autor indica como fases ou momentos do processo, mais especificamente, o reconhecimento, a redução e a representação. A seguir são explicitados os referenciais semânticos que nos levam à compreensão do que se entende por texto e por documento e à superação da dicotomia significado/significante pela união dos planos sintático, semântico e pragmático no discurso, considerado enquanto seqüência de microestruturas. A organização dos textos em macroestruturas que representam seu significado global é explorada ainda nesse primeiro capítulo, que finaliza com uma análise da superestrutura e identificação das partes do texto, em que são ilustrados dois esquemas típicos, o da narração e o da investigação.

No segundo capítulo da obra, sobre o reconhecimento ou leitura do documento, o autor nos apresenta aspectos identificadores da leitura com finalidade analítico-documentária, tendo presentes inicialmente os processos inferenciais nesse processo, que nos permitem comparar o que já sabemos com o trazido pelo texto. Refere-se às inferências *elaborativas*, ou projeções de nossos esquemas cognitivos no texto, às inferências *redutivas*, que nos permitem identificar o essencial na mensagem e, paralelamente, por meio das inferências *lógico-sintáticas,* compreender como está construído o texto, assim como, apoiados nas inferências *léxicas,* captar informações a partir dos conceitos que as palavras expressam.

São então indicadas estratégias para a leitura dos textos, segundo os dois estágios no processo de reconhecimento do documento, a leitura de situação e a leitura ativa. Quando o processo de análise é realizado por pessoas, o exame inteligente do texto enfoca os lugares mais ricos para se obterem informações, conforme recomendação de Anderson, podendo ser regido por questões chaves como as de Lasswell ou os critérios de Cícero, em *De oratore*, de algum modo representados na gramática de casos de Filmore e correspondentes às facetas do sempre atual método de Ranganathan para análise do conteúdo dos documentos. A seguir, finalizando o segundo capítulo, o autor nos apresenta recomendações para a redução do texto, indicando táticas, considerações e exemplos de critérios suscetíveis de aplicação prática.

O processo de indexação e seus resultados, os índices, são objetos do terceiro capítulo, cuja primeira parte trata do conceito de indexação e de seu procedimento. Na segunda parte, sobre critérios e condições para uma boa indexação, são descritos os objetivos que o analista deve perseguir, tais como a especificidade, a relevância, a exaustividade e a precisão, destacando-se a entropia, a procedência dos termos, a profundidade, o índice de consistência e outros indicadores de avaliação. A seguir, são enfocadas a seleção e a atribuição dos termos,

tarefas inspiradas nas necessidades dos usuários e principalmente fundamentadas no contexto das culturas às quais pertencem e nas suas experiências pessoais. Analisa ainda elementos do universo de possibilidades para representação dos conceitos selecionados, desde vocabulários controlados às linguagens livres e às possíveis circunstâncias que levam à decisão de se incluir um termo como representante do conteúdo original.

Entre tais circunstâncias, encontram-se as determinantes dos níveis de indexação, ordenados na quarta parte do capítulo, conforme sua aspiração a uma proposta mais genérica ou mais seletiva: classificação ou categorização; indexação superficial; indexação profunda; indexação exaustiva e indexação seletiva. O autor nos oferece uma extensa reflexão sobre os índices, sua natureza e categorização, na quinta parte do terceiro capítulo, permitindo que cheguemos à compreensão do universo categorial de abrangência do conceito índice. Trata-se de uma exposição didaticamente irrepreensível, pela qual ele concretiza na obra seus conhecimentos sobre o tema, aplicando metodologicamente o que demonstra na teoria. Distingue inicialmente os índices livres, baseados em palavras do texto dos índices controlados, baseados em conceitos. Entre os primeiros, inclui os índices de documentos individuais (nomes próprios, geográficos, topográficos e cronológicos), os de coleções de documentos, entre os quais destaca os índices esquemáticos, os índices de palavras e nomes, os permutados (tipos KWIC, KWOC e KWAC), os índices de unitermos e os índices de citações. Por sua vez, os índices baseados em conceitos abrangem os índices analíticos de livros, revistas e bibliografias, os índices classificados, os sistemas de índices coordenados de recuperação da informação mediante operadores lógicos e os boletins de índices sistemáticos. A quinta parte do terceiro capítulo é destinada à relação índices e Internet, analisando questões sobre a indexação com motores de busca, tanto no que se refere à recuperação por palavras-chave, aos metadados e indexação de documentos digitais, quanto à recuperação conceitual na Internet. A parte final do capítulo é dedicada à indexação automática e, ao iniciar suas reflexões sobre o tema, o autor a distingue da indexação assistida por computador e da indexação semi-automática. Analisa então os modelos extrativos de caráter estatístico e probabilístico, cuja origem coincide com as primeiras tentativas de conjugar a informática e a estatística à documentação. A essência do processo é a identificação automática de palavras-chave no texto pela freqüência com que aparecem e sua fundamentação teórica tem origem na lei de Zipf. Novas formulações desta Lei originaram outras técnicas de discriminação dos termos, sobre as quais discorre o autor, destacando a indexação estatística de termos por freqüência, conhecida pela sigla IDF, a *Term frequency, inverse document frequency* (TFIDF), o método *N-grams*, que modifica a lei de Zipf para possibilitar o tratamento de palavras compostas e os *Stemmers*, que utilizam a freqüência com que aparecem seqüências

de letras no corpo de um texto para extrair a raiz das palavras. Além dessas possibilidades, as relações semânticas entre os termos lingüísticos podem ser estabelecidas por métodos de agrupamento e classificação.

Ainda relacionados à indexação automática, os modelos analíticos de caráter lingüístico são derivados do processamento da linguagem natural a aplicados desde os anos 60, sob o impacto das teorias lingüísticas e fundamentados em processos analíticos de natureza morfológico-léxica, sintática, semântica ou pragmática. A seguir o autor nos apresenta procedimentos e critérios em prol de um processamento inteligente e, para finalizar o capítulo 3, nos descreve alguns programas de indexação automática que combinam o modelo lingüístico com ferramentas estatísticas.

As linguagens que representam o conteúdo dos documentos são o tema do quarto capítulo, que procura oferecer um panorama abrangente da variedade e da evolução histórica dessas linguagens, para cujo estudo recomenda-se que se atente, por um lado, às considerações de ordem lingüística e, por outro, às condições funcionais e ferramentas precisas, a serem utilizadas em contextos e necessidades determinados e, por sua vez, também determinantes.

Na primeira parte do capítulo são analisadas as linguagens naturais, distintas inicialmente em suas modalidades geral e científica e em sua utilização documental, que se verifica de forma livre ou controlada.

Quanto à tipologia das linguagens documentais, objeto da segunda parte do capítulo, o autor no apresenta: a linguagem livre, representada por listas de unitermos, listas de palavras-chave e glossários; as linguagens controladas (pós-coordenadas), representadas pelas listas de cabeçalhos de assunto e tesauros e as linguagens codificadas (pré-coordenadas), ou sistemas de classificação. A seguir, discute a indexação mediante linguagem livre e linguagens controladas, apresentando suas características, vantagens e desvantagens, para então discorrer mais especificamente sobre a informação representada mediante tesauros. Nesta sessão do trabalho são tratadas as relações terminológicas, tais como de equivalência, definitórias, hierárquicas ou classificatórias e de associação; as fases típicas na construção de um tesauro, desde a terminológica, passando pela fase documental, pelas formas de apresentação hierárquica e alfabética, até a elaboração de índices e fase de difusão. Trata ainda da superestrutura do documento tesauro, quando descreve o plano global de sua apresentação, das tendências na sua construção, dos tesauros em linha e dos mapas conceituais de redes semânticas, método de representar o conhecimento no campo da inteligência artificial, para complementar a função comunicativa da linguagem. Ao destacar os *topic maps*, o autor apresenta suas possibilidades de proporcionar acesso à informação existente em diferentes

redes semânticas, embora não menospreze os limites às possíveis aplicações desse novo paradigma, incluindo no capítulo um ilustrativo quadro em que são apresentadas as relações entre as características do mapa conceitual, dos *topic maps* e dos tesauros. A seguir, apresenta um modelo de aplicação baseado nas exigências para a criação e gestão automática de tesauros, discorre sobre a geração automática de tesauros de verbos, seus fins, aplicações e modalidades de organização e sobre a representação codificada da informação. Finaliza o capítulo dissertando sobre outros esquemas de representação da informação, com destaque para as ontologias representativas do conhecimento em inteligência artificial e sua representação por meio das técnicas de engenharia de software.

O quinto e último capítulo é dedicado ao resumo científico, iniciando-se com a natureza e finalidades desse tipo de resumo. São enumeradas, a seguir, as regras básicas da representação, embora o autor reconheça que o autor do resumo não é somente um intermediador, mas também um criador, cuja tarefa transcende as cadências pré-estabelecidas. Assim, denomina de valores as considerações que pautam a construção e redação do resumo: entropia; pertinência, coerência; correção lingüística ou correção gramatical e estilo. Discorre ainda sobre alguns modelos de resumos para, então, discutir sobre o processamento automático, desde os primeiros métodos extrativos, passando pelos modelos lingüísticos e cognitivos, até a síntese dos documentos múltiplos. Encerra o capítulo analisando os critérios para avaliar a elaboração dos resumos, basicamente o grau de reutilização, o traslado da superestrutura do original, a qualidade técnica, o tamanho e densidade e a coesão dos resumos.

Como resultado da leitura, pode-se concluir que os ensinamentos transmitidos apresentam-se num contexto teórico exaustivamente analisado e discutido, em que se realizam aproximações atuais relativas à análise documental especialmente no que se refere à abrangência temática, à qualidade, atualidade e uso crítico do quadro referencial.

Fartamente ilustrado, o texto consegue o equilíbrio necessário para se tornar ao mesmo tempo profundo e interessante. Além de apresentar o estado da questão sob uma perspectiva lingüística, o autor atende a preocupação dos indexadores em estabelecer relações entre a linguagem natural e as linguagens documentais. Assim, pode-se afirmar que a obra é bem vinda e justifica congratulações ao autor pela erudição e o estilo que demonstra, qualidades que se equilibram numa obra de valor, tanto do ponto de vista estrutural, quanto gráfico.

O livro faz parte de uma coleção de títulos voltados à área de Ciência da Informação, todos de indiscutível qualidade gráfica, fato que revela a consistência da produção editorial da Ediciones TREA, de Gijón (Astúrias). Com o intuito de facilitar a obtenção da obra pelos

leitores, seu representante pode ser encontrado no endereço jmpujol@pujolamado.com, de Pujol y Amado, coordenador para a América.

Originais recebidos em 24/11/2004.