CDU 025.4.03

ANÁLISE DE ASSUNTO

Maria Augusta da Nóbrega CESARINO, Professora da Escola de Biblioteconomia da Universidade Federal de Minas Gerais.
Maria Cristina Mello Ferreira PINTO, Professora da Escola de Biblioteconomia da Universidade Federal de Minas Gerais.

A análise de assunto como operação-base para a recuperação da informação. Diferentes níveis de análise de assunto: a) para determinar o conteúdo informativo de um documento; b) para compreender a necessidade de informação transmitida pelo usuário; c) para escolher ou criar a linguagem de indexação mais adequada para o sistema. Recomendações a serem observadas em cada uma dessas situações. A análise de assunto automática.

1. INTRODUÇÃO

A análise de assunto é a operação-base para todo procedimento de recuperação de informações. Normalmente são estas as situações nas quais os bibliotecários fazem análise de assunto:

a) quando recebem um documento e devem dar entrada deste num sistema de informações. Nesta situação farão uma análise com o objetivo de determinar o conteúdo informativo do documento em questão, tendo em vista o objetivo do sistema e as necessidades dos usuários;

b) ao receberem um pedido de informação, devem fazer uma análise deste com o objetivo de compreender a necessidade de informação transmitida pelo usuário, identificar os conceitos existentes no pedido e traduzí-los para a linguagem adotada pelo sistema.

Esta duas situações estão bem definidas no diagrama de LANCASTER (5), representado na Figura 1. Nele percebemos que, em qualquer sistema de recuperação de informações, a análise de assunto está presente não só no momento da indexação, como também na etapa de busca da informação.

Enfim, outra situação ocorre ainda quando o bibliotecário é levado a escolher o melhor modo de indexar uma coleção de documentos. Ele terá de analisar o assunto da coleção como um todo, isto é, a área do conhecimento humano coberta pelo sistema. Este procedimento é bem mais complexo, considerando-se a gama de variáveis que poderão interferir na escolha.

2. ANÁLISE DE DOCUMENTOS

Ao darem entrada em qualquer sistema de recuperação de informações, os documentos serão analisados de duas maneiras:

a) bibliografica ou objetivamente: este tipo de análise pretende a descrição do documento em termos de suas características físicas, respondendo à questão: "Qual a aparência deste documento?";

b) intelectual ou subjetivamente: este tipo de análise pretende a descrição do

documento em termos de seu conteúdo informativo (assunto abordado), respondendo à questão: "De que trata este documento?".

Se um documento é processado em diferentes locais, em pólos opostos de um país, ou mesmo do mundo, a descrição bibliográfica ou objetiva deve resultar idêntica (pelo menos no que se refere aos dados essenciais) e a descrição intelectual ou subjetiva poderá variar. Isso porque, para a descrição física dos documentos, nós nos baseamos em elementos que são facilmente identificáveis e que por si caracterizam realmente a obra.

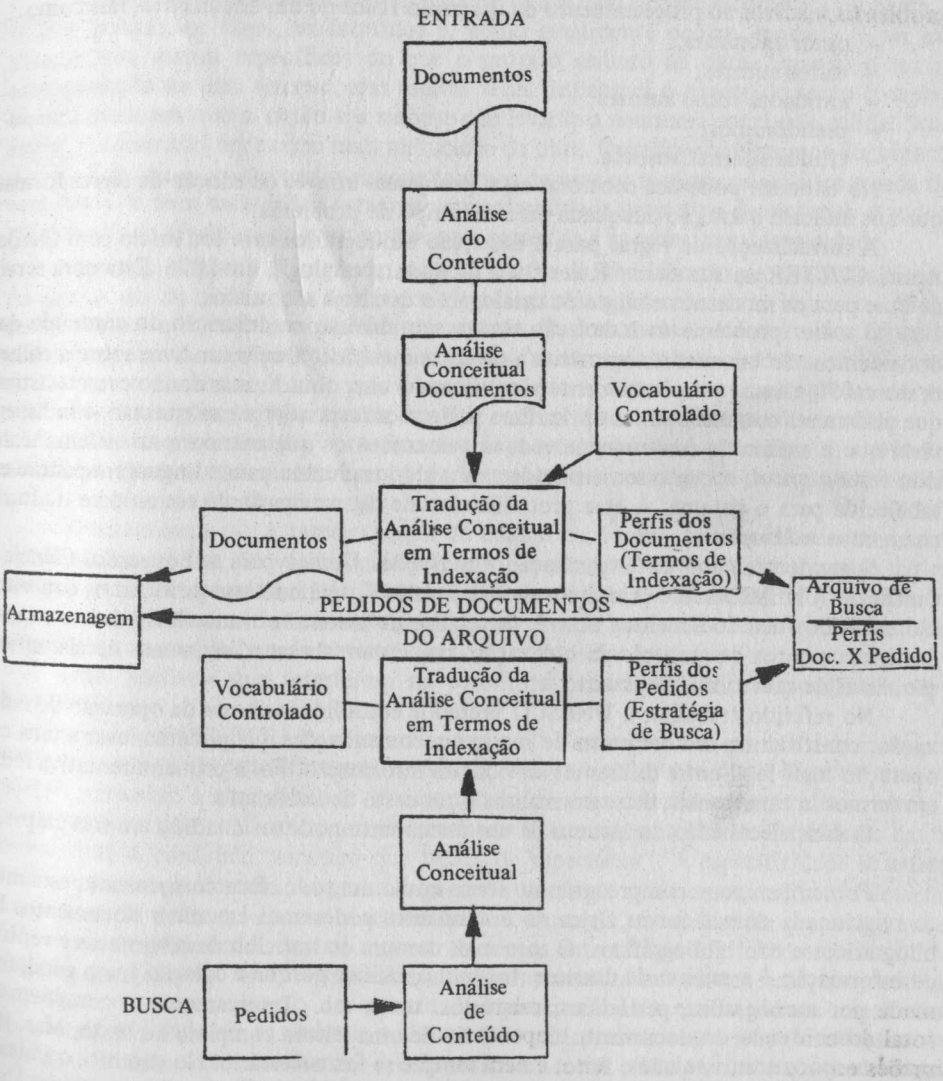

Fig. 1 Recuperação da informação: Processo de entrada e saída

Temos, como exemplo:

- o número de especificação de uma patente;
- o nome do autor da obra;
- o título do documento;
- a edição, o editor, a data de publicação, etc.

Estes são alguns dos chamados elementos identificadores de uma obra. Para se produzir uma descrição única de um documento é necessário se descrever suas características físicas, usando um conjunto ordenado de dados bibliográficos. É certo que existem problemas relativos ao processamento da descrição física de um documento; tais como:

- obras anônimas;
- vários autores;
- entidades como autores;
- pseudônimos;
- títulos alternativos, etc.

No entanto, podemos controlar esses problemas através da adoção de regras formais que nos indicam a solução adequada para cada tipo de problema.

A formalização de regras para a descrição bibliográfica teve seu início com Charles Ammi CUTTER na sua obra "Rules for a dictionary catalog", em 1876. Esta obra serviu de base para os modernos códigos de catalogação que hoje são usados.

O maior problema da indexação reside, sem dúvida, na descrição do conteúdo dos documentos. Se um usuário apresenta a seguinte questão: "Quero um livro sobre a colheita do café", ele não estará realmente identificando uma obra X, mas dando características que podem ser comuns a várias obras. Para podermos responder a essa questão, é indispensável que a análise de conteúdo de todos os documentos que entraram no sistema tenha sido feita *a priori*, e que o seu resultado tenha sido traduzido para a linguagem-padrão estabelecida para o sistema. A esse procedimento de determinação de conteúdo e tradução chamamos *indexação*.

Segundo trabalho da Organização das Nações Unidas para a Educação, Ciência e Cultura – UNESCO sobre princípios de indexação (1), durante esse processo os conceitos são extraídos dos documentos através de análise de assunto e transcritos para os termos dos instrumentos de controle da indexação, tais como: thesauri, esquemas de classsificação, listas de cabeçalhos de assunto, etc.

No referido trabalho, a UNESCO pretende consolidar as bases da operação de indexação, constituindo um conjunto de regras e recomendações que poderão levar a uma cooperação mais fácil entre diferentes serviços de informação. Foi a primeira tentativa feita, em termos internacionais, de se normalizar o processo de indexação.

O estabelecimento do assunto de um documento pode ser dividido em três etapas, a saber:

Primeira etapa – compreensão do texto como um todo. Esta compreensão está muito relacionada com a forma física do documento, poderemos examinar documentos bibliográficos e não bibliográficos. O caso mais comum do trabalho de bibliotecas e centros de informação é a análise de documentos bibliográficos porque a coleção é, em geral, formada por monografias, periódicos, relatórios, teses, etc. Teoricamente, a compreensão total do conteúdo do documento dependeria de uma leitura completa do texto. Mas, por razões econômicas, isso não é feito, e nem sempre se faz necessário. No entanto, o indexador deve ser certificar que nenhuma informação necessária passou-lhe despercebida. Uma atenção particular deve ser dada a partes importantes do texto:

- título;
- introdução e subtítulos dos capítulos e parágrafos;
- ilusrações, tabelas, diagrmas;
- conclusões;
- palavras ou grupos de palavras que tenham sido graficamente diferenciadas: sublinhadas, em cursivo, etc.

As intenções do autor são geralmente relatadas nas seções introdutórias, enquanto que as seções finais mostram o quanto suas proposições alcançaram do objetivo proposto.

A indexação baseada exclusivamente no título nem sempre é completa ou exata. Os títulos podem ser vagos, inadequados e, como geralmente ocorre, tendem a ser mais abrangentes, menos específicos do que o próprio assunto da obra. Quando o texto é acompanhado de uma sinopse, esta muitas vezes dispensará o exame do texto completo. Haverá casos em que o título e a sinopse não levarão a nenhuma conclusão válida. Sendo assim, é necessário um exame mais minucioso da obra. Quando examinarmos documentos não bibliográficos, como audio-visuais, teremos de usar os recursos adequados a cada tipo específico. É bem mais difícil o exame em profundidade desse tipo de material. A indexação a partir do título ou da sinopse, no caso, tornar-se-á às vezes, uma necessidade.

Segunda etapa – A segunda etapa é relativa à identificação de conceitos. Depois do exame do documento, o indexador deve seguir um procedimento lógico na seleção dos conceitos que melhor expressarão o assunto do documento. A escolha dos conceitos poderá espelhar a configuração das categorias fundamentais reconhecidas como importantes para o assunto abordado pelo documento, isto é, objetos, materiais, processos, propriedades, operações, equipamentos, etc. Por exemplo: ao examinar um documento sobre "aplicação de quimioterapia" o indexador deve examinar sistematicamente o texto no sentido de listar outros elementos abordados pelo mesmo: o tipo de doença, efeitos colaterais, composição das drogas, órgão tratado, etc.

Terceira etapa – A terceira etapa do exame do documento é a da seleção de conceitos que realmente são válidos para serem indexados. Nem sempre é necessário dar entrada a todos os conceitos identificados durante a análise do documento. A escolha dos conceitos dependerá exclusivamente do objetivo para o qual o documento está sendo indexado, isto é os objetivos do sistema e as necessidades dos usuários.

Duas variáveis que interferem na escolha de conceitos são: a especificidade e a exaustividade.

Segundo FOSKETT (2), a especificidade é " a extensão em que o sistema nos permite ser precisos ao especificarmos o assunto de um documento que estejamos processando" enquanto a exaustividade "é o resultado de uma decisão administrativa, sendo a extensão com que analisamos um determinado documento, a fim de estabelecer exatamente qual o conteúdo temático que temos de especificar". A especificidade se refere, então, ao estabelecimento do grau de precisão com que poderemos realmente determinar o assunto principal de um documento. Por exemplo, na análise de um artigo sobre "Medidas de prevenção contra sequestros aéreos em aeroportos internacionais", um sistema que admita uma especificidade maior permitirá selecionar este conjunto de conceitos para representação do assunto: "Sequestros". Já a exaustividade relaciona-se com a possiblidade de se indexar o documento em profundidade *(depth indexing)* na medida em que o sistema admita a indexação do tema principal do documento como também de sub-temas. Em bibliotecas mais gerais, o grau de exaustividade é sempre menor do que o necessário para bibliotecas especializadas.

Considerando-se o exemplo anterior, um sistema com grande grau de exaustividade permitiria que se desse entrada para "Aeroporto de Orly", "Aeroporto do Galeão"..., se o documento tratasse desses aeroportos específicos como sub-tema.

Estes dois fatores, especificidade e exaustividade, influenciam não somente a seleção dos conceitos para um determinado documento, como têm influência praticamente em todo processo de recuperação da informação. MEADOW (3) mostra a ligação desses fatores com as medidas de revocação e precisão que estabelecem a eficiência dos sistemas de recuperação da informação. Através de pesquisas sobre o desempenho de vários sistemas, ele demonstrou que um maior grau de especificidade eleva a taxa de precisão e baixa a de revocação: ao contrário, um aumento de exaustividade eleva a taxa de revocação, baixando a de precisão.

Depois de estabelecido o assunto de um documento, o passo seguinte será a transformação dos conceitos selecionados em termos ou símbolos autorizados para representá-los no sistema. Para isso faremos uso de instrumentos de controle de linguagem, tais como: thesauri, listas de cabeçalhos de assunto, sistemas de classificação, etc.

O indexador deverá estar familiarizado com a linguagem-padrão do sistema em que trabalha. Toda linguagem é formada de vocabulário e sintaxe. O vocabulário compõe-se de unidades isoladas enquanto a sintaxe estabelece o modo pelo qual essas unidades serão combinadas para a efetiva comunicação do pensamento. Assim o indexador deverá dominar tanto o vocabulário como a sintaxe do seu instrumento de indexação.

Embora seja importante ressaltar que a existência de uma linguagem-padrão no sistema de recuperação de informações não deve influenciar a análise conceitual dos documentos, muitas vezes esta linguagem impõe limitações à representação do assunto. Por exemplo: o esquema de classificação usado não nos permite ser tão específicos quanto necessitaríamos para indexação do documento sobre "medidas de prevenção contra sequestros aéreos em aeroportos internacionais". Se o instrumento usado é um thesaurus ou uma lista de cabeçalhos de assunto, o número de termos fixados para representar o conteúdo do documento poderá ser limitado, diminuindo consequentemente o número de entradas, isso porque, nesses vocabulários, são estabelecidas, *a priori*, as relações entre os termos. Bastará que o documento seja indexado pelo descritor mais específico. Se o sistema usado é um esquema de classificação, o indexador deverá estar familiarizado com os simbolos, regras para sua formação, etc.

Na prática, o indexador freqüentemente encontrará conceitos que não estão representados na linguagem usada. Dependendo do sistema, a admissão dos novos termos é aceita; em outros casos o indexador deverá usar descritores mais genéricos.

Um outro aspecto a ser considerado na indexação de documentos é o controle de qualidade da indexação. Esta consistência da indexação está ligada a dois elementos básicos: ao desempenho do indexador e à qualidade dos instrumentos de indexação. Para a operação eficiente de qualquer sistema de informação é importante conseguirmos uma boa consistência na indexação independentemente do ponto de vista do indexador. É importante também que esta consistência seja regular, considerando-se o fator tempo na operação de um determinado sistema, sendo necessário ao indexador um alto grau de imparcialidade e uma submissão às diretrizes da indexação adotadas pelo sistema. Mas é quase inevitável que elementos de subjetividade interfiram na indexação. No entanto, devem ser minimizados tanto quanto possível. A consistência é mais difícil de ser conseguida quando o grupo de indexadores é grande, ou quando trabalham em diferentes locais. Nesses casos é recomendável que se estabeleça um grupo de controle centralizado para a verificação das indexações feitas.

O *background* do indexador ainda é um ponto em discussão. Ele necessariamente deve ser um especialista na área coberta pelo sistema? Ou a eficácia de seu desempenho estará mais relacionada com a sua experiência em indexação, preferivelmente na área, do que à sua formação acadêmica?

O projeto Cranfield I, testando três indexadores, conclui que o resultado da indexação não era afetado pela formação acadêmica deles mas pela experiência prévia que possuiam (9).

LANCASTER sugere que o vocabulário controlado é indispensável nas etapas de indexação e pesquisa dos sistemas de recuperação não-mecanizados, enquanto que experiências consideráveis em sistemas automatizados mostram que estes operam, de maneira eficiente, usando a linguagem natural dos documentos. Nestes casos, a presença de um elemento humano de alto nível é muito mais necessária na etapa de pesquisa do que no momento da indexação. A complexidade da análise de assunto está no entendimento da questão do usuário e na montagem da estratégia de busca. É indiscutível que o controle de qualidade da indexação poderá lucrar muito através do estabelecimento de um contato mais efetivo entre o indexador e o usuário.

A consistência da indexação será também muito afetada pela qualidade da linguagem adotada. O estabelecimento da linguagens de indexação deverá ser feito tendo em vista a sua adequadação ao sistema a que irá servir. Um fator importante é a qualidade de atualização dessas linguagens em resposta a novos desenvolvimentos na terminologia, necessidades dos usuários e do próprio sistema.

SLAMECKA citado por LANCASTER (5) diz que a linguagem controlada pode exercer duas funções: a prescritiva ou a susgestiva. A prescritiva estabelece limites rígidos para a representação dos conceitos e a sugestiva, sendo mais flexível, indica as melhores formas de representação, sem impô-las ao indexador. A linguagem prescritiva facilita a escolha dos termos e, por conseguinte, a consistência da indexação permitindo menos interferência pessoal do indexador. Por exemplo: o thesaurus é mais sugestivo do que prescritivo, ao contrário das listas de cabeçalho de assunto.

O procedimento da análise de questões propostas pelos usuários é semelhante ao procedimento da análise de asuntos de documentos em sistemas de recuperação da informação. Como no caso da análise de documentos, teremos de analisar as perguntas dos usuários, passando por todas as fases aqui descritas. Teremos de compreender a necessidade do usuário como um todo, identificar os conceitos que nela estão implícitos, selecionar aqueles que são válidos e transformá-los em termos ou símbolos usados no sistema de indexação adotado. Nessa análise de questões um fator importante é a interação usuário x indexador x sistemas. Os sistemas que permitem este tipo de interação tem demonstrado ser mais eficientes do que aqueles onde ela não existe. Segundo HATT (4) "É possível que as grandes dificuldades encontradas por usuários de bibliotecas, catálogos e outros isntrumentos bibliográficos sejam devidas ao fato de que procuram encontrar documentos através de instrumentos desenvolvidos para reencontrá-los ou recuperá-los *(retrieval)*. Quando um classificador classifica uma obra, ele está determinando um lugar para aquele texto numa estrutura do conhecimento, da qual ele tem boa noção. O usuário procurando aquele mesmo texto está numa posição muito diferente". Os termos ou categorias que são estabelecidos quando estamos estruturando um campo do conhecimento não são os mesmos termos e categorias usados quando estamos procurando um ítem de informação. No primeiro caso podemos e devemos definir uma categoria com relação a outras existentes no campo. No segundo caso pode não existir uma visão clara do campo, muito menos a visão

estruturada que irá produzir categorias nas quais encaixaremos um determinado ítem de informação.

3. ANÁLISE DE COLEÇÕES DE DOCUMENTOS

Outra situação em que o bibliotecário normalmente faz análise de assunto é quando ele examina uma coleção de documentos e a área de assunto coberta por essa coleção, com o objetivo de escolher ou mesmo criar uma linguagem ideal para a sua indexação.

Os primeiros estudos visando à classificação de áreas do conhecimento dividiam-nas do geral para o específico. É o caso das tradicionais classificações bibliográficas, também chamadas de hierárquicas, que criavam cadeias de termos seguindo o princípio de hierarquia.

Foi a partir das proposições feitas por RANGANATHAN que a análise das áreas do conhecimento passou a ser feita inversamente, ou seja, partindo do específico para o geral, baseando-se em categorias fundamentais. Os menores ítens de informação de um campo do conhecimento eram agrupados segundo alguma semelhança formando então as categorias fundamentais para o assunto. RANGANATHAN desenvolveu sua "Colon Classification", que é uma classificação geral, dividindo o conhecimento humano com base nas categorias fundamentais (Personalidade, Matéria, Energia, Espaço, Tempo – PMEST) aplicáveis a todos os campos do conhecimento. Baseadas nas idéias de RANGANATHAN, fizeram-se várias pesquisas, resultando em classificações chamadas facetadas, para inúmeras áreas especializadas do conhecimento, sendo que não mais se tenta adaptar as categorias de RANGANATHAN (PMEST) para cada área, mas criam-se categorias específicas de cada uma, de acordo com as necessidades próprias.

Embora a análise facetada tinha sido comumente usada para a criação de sistemas de classificação, pode ser aplicada, com sucesso, à elaboração de qualquer tipo de vocabulário controlado. Como a matéria prima para a análise facetada são os próprios termos que representam os conceitos da área, pode-se tentar várias maneiras de levantar esses termos.

LANCASTER (5), apresenta de uma forma muito clara quatro abordagens para se gerar um vocabulário controlado:

1) gerar um vocabulário empiricamente com base na endexação de um conjunto representativo de documentos;

2) modificar um vocabulário já existente, como por exemplo, transformar uma lista de cabeçalho de assunto em thesaurus;

3) extrair o vocabulário de um outro já existente. Por exemplo, do vocabulário usado em uma área geral cria-se um micro-vocabulário de uma área específica, subordinada:

4) reunir termos de diferentes fontes: especialistas na área, dicionários, glossários, índices, etc.

Ainda segundo o autor, o 1º e o 4º métodos são normalmente os mais usados, enquanto que o 2º e o 3º são usados somente dentro de certas condições.

No primeiro método, a partir da livre indexação de um conjunto de documentos, teremos um vocabulário inicial. Os termos deste vocabulário base serão agrupados segundo características comuns, analisados, selecionados e estruturados de uma maneira lógica. Este método foi usado pela Dupont de Nemours para elaboração de seu thesaurus, que lançou as bases para os modernos thesauri usados hoje em várias áreas do conhecimento. Este método é chamado de empírico ou analítico. No quarto método, chamado formal *(committee approach)*, levantamos os termos com base em entrevistas com especialistas na

área. A lista de termos é então analisada por um corpo editorial. Somente quando o vocabulário está na sua forma final ou semifinal é que é usado para indexar documentos como teste. Esse método foi adotado pelo Enginnering Joint Council na compilação do Thesaurus of Engineering Terms.

A construção do Thesaurus of Engineering and Scientific Terms (TEST), gerado segundo o método formal, obedeceu a alguns critérios na seleção dos termos:

- aceitabilidade do termo em dicionários, enciclopédias, etc;
- utilidade do termo em comunicações, em índices e em sistemas de recuperação da informação;
- o número de fontes que usam este termo;
- a clareza e precisão do termo;
- a pertinência desse termo com outros já selecionados.

A elaboração do Thesaurus of DDD Descriptors (ex-Thesaurus of ASTIA Descriptors) baseou-se numa lista de cabeçalho de assunto, modificando a sua estrutura e reduzindo os seus termos de modo a construir um novo vocabulário.

Ainda que especialistas no assunto sejam de grande utilidade na compilação de linguagens de recuperação da informação, seria desvantajosos nos limitarmos somente a eles como base para análise e estruturação de uma área. As desvantagens apontadas por LANCASTER, quando nos baseamos somente em especialistas, são :

- podem não estar inteiramente familiarizados com a literatura, e mais importante, com as necessidades dos usuários potenciais do sistema;
- podem tomar decisões que não são úteis tendo em vista o objetivo de recuperar informações;
- podem dar mais importância à sua própria especialidade, causando desequilíbrio no sistema como um todo.

Um vocabulário controlado é essencialmente um instrumento prático, devendo ser capaz de representar conceitos que realmente ocorrem na literatura do assunto. Torna-se óbvio que terá de ser então coerente com a mesma. Este é o princípio da "garantia literária" como foi chamado por HULME – "Principles of Library Classification" (1911) – defendendo a idéia de que as "classes" usadas para agrupar documentos não deveriam ser baseadas em nenhuma classificação teórica do conhecimento, mas sim nos grupos que os documentos parecem formar logicamente por si mesmos, isto é, classes sobre as quais a literatura existe".

Um vocabulário desenvolvido empiricamente a partir da indexação da literatura da área tem grande garantia literária, enquanto que um vocabulário desenvolvido por especialistas pode não ter nenhuma.

Um dos erros na construção de vocabulários controlados, quer usando o método formal ou o empírico, tem sido a omissão do estudo e aproveitamento das questões propostas pelos usuários. Não basta a garantia literária, porque podem existir termos usados com freqüência na literatura e que pouco aparecem nas pesquisas dos usuários. Esta nova aboradagem propõe que, combinada com o levantamento dos termos na literatura, se faça também a análise de perguntas feitas aos sistemas de recuperação da informação da área.

Depois de se levantarem os termos mais importantes na área, o passo seguinte é o agrupamento destes em facetas homogêneas e mutuamente exclusivas. Nesta etapa, a análise facetada pode ser utilizada com vantagens.

Podemos resumir as vantagens da análises dizendo que tal processo ajuda a:

- determinar as principais categorias de termos para uma área;
- optar pelo melhor termo para representar um conceito;
- estabelecer quais as relações úteis entre os termos tendo em vista a recuperação de informações;
- estabelecer hierarquias necessárias.

A análise facetada de qualquer área do conhecimento evidência as relações existentes entre termos e categorias. Para o indexador o conhecimento dessas relações é de suma importância, considerando a necessidade cada vez mais comum de indexarmos assuntos compostos e complexos. Devido à especialiação das ciências, hoje em dia é muito mais provável recebermos um documento sobre "Barulho de turbinas de aviões", do que sobre "Aviões". Daí a necessidade de conhecermos as relações entre os termos e categorias, para podermos prever nas nossas linguagens de indexação a possiblidade de representações adequadas para assuntos compostos e complexos, ou seja, adequadas para o entrelaçamento dos assuntos. Podemos dividir em três tipos as relações normalmente existentes entre termos:

- relações de equivalência;
- relações hierárquicas;
- relações associativas.

Relações de equivalência (ou relações preferenciais). – Alguns conceitos podem ser representados por mais de um termo (termos similares ou de significação quase idêntica) como: revista e periódico. O indexador, quando analisa um documento cujo assunto pode ser representado por diversos termos equivalentes, deve escolher apenas um termo para representar o conceito. Essa opção pode ser orientada para:

- o termo mais conhecido;
- o termo mais claro ou que se presta menos a ambiqüidade.

O indexador não deve esquecer que os outros termos também devem estar representados no índice.

Exemplo: optar entre:

- alteração ou modificação;
- curvatura ou flexão;
- genética ou hereditariedade;
- feedback ou retroalimentacão.

Relações hierárquicas.– Expressam uma idéia de subordinação entre os termos. Dentre os tipos mais comuns de relações hierárquicas, podemos indicar:

- relação gênero-espécie:

Exemplo:

processamento técnico
  registro
  catalogação
  classificação
  indexação

- relação todo/parte:

Exemplo:

árvore
  raiz
  tronco

galhos
folhas

Na indexação, a preferência é dada aos assuntos específicos. Por exemplo, dentro de uma coleção que trata de esportes, os documentos que falam especificamente de basquete devem ser indexados neste termo. Entretanto, numa coleção geral, com poucos documentos sobre esportes, para evitar dispersão, eles podem ficar agrupados sob esse termo.

Relações associativas (ou relações de coordenacão). — Alguns conceitos estão estreitamente ligados uns aos outros, de modo que a idéia de um faz lembrar a idéia do outro. Existem vários tipos de relações associativas:

- genética: pais e filhos
- causa e efeito: ensino e aprendizagem
- instrumental: escrita e lápis
- material: papel e livro
- similaridade de processo: catalogação e classificação

As relações de coordenação só devem ser mantidas na indexação quando facilitam realmente a recuperação da informação.

Exemplo:

Analisando-se o livro "Classificação e Indexação em Ciências Sociais", dentro da biblioteca de um Curso de Bibliotecnomia, não seria correto indixar os dois temos: Classificação e Indexação juntos, embora haja uma similaridade de processos entre eles. A indexação do mesmo documento numa biblioteca de Ciências Sociais poderia permitir tal prática. Isto porque, no primeiro caso, os dois assuntos são importantes para os usuários e há documentos em quantidade para cada um deles. No segundo caso os assuntos são de menor importância para o usuário.

Esses três tipos de relações, existentes em qualquer área, devem ser evidenciados nas linguagens controladas através dos recursos prescritivos ou sugestivos: é a utilização, nas listas de cabeçalhos de assunto, remissivas *ver* e *ver também* e nos thesauri das relações TG (termo geral), TE (termo específico) TR (termo relacionado) e USE. Deve-se procurar, sempre que possível, mostrar todas as relações entre os termos: relações ascendentes, descendentes e colaterais.

Outro problema a se enfrentrar, ao criar uma linguagem de indexação, é o estabelecimento de uma ordem de prioridade para os diversos conceitos. Essa ordem é determinada a partir da importância que os conceitos têm para os usuários. Algumas linguagens de indexação exigem que o indexador pré-determine a ordem de citação dos conceitos. São as chamadas linguagens pré-coordenadas (cabeçalhos de assunto, sistema de classificação). Algumas vantagens de se pré-determinar a coordenação de conceitos são: prover um e apenas um lugar inequívoco para qualquer assunto composto e fazer com que os usuários possam familiarizar-se com o sistema e, com o tempo, passem a formular questões de modo que se adaptem a eles.

Vários autores criaram regras básicas tentando estabelecer a ordem de citação dos assuntos complexos.

Exemplo:

- o assunto antes da forma bibliográfica - Física - Dicionário;
- o assunto antes do lugar (com a preposição em) - "A educação no Brasil" - Educação-Brasil;

– o assunto depois do lugar (com a preposição de) - Rios do Brasil - seria Brasil-Rios.
– concreto, depois o processo - Tratamento de metais seria Metais-Tratamento.
– todo-parte – índices de Revistas seria Revistas-Índices.

Qualquer que seja a ordem de citação escolhida, temos de aceitar o fato de que não podemos sempre agradar a todo mundo. As linguagens pós-coordenadas tentam eliminar essa deficiência, isolando conceitos e deixando que o usuário relacione os coneitos da forma que quiser, no momento da busca.

4. AUTOMAÇÃO E ANÁLISE DE ASSUNTO

Com o desenvolvimento de equipamentos de processamento de dados, é natural que os computadores sejam aplicados ao controle da informação. Pesquisas têm estudado a possibilidade de análise automática de textos, tarefa até então executada exclusivamente por indexadores.

A primeira idéia de se usar computadores no processamento lógico da análise de assunto de documentos foi proposta por LUHN (7), em 1959. Indicava o uso de títulos dos documentos como "matéria-prima" para representação de assunto. Nesse caso, não havia propriamente *análise de assunto* porque os documentos não eram analisados, com também não havia controle de vocabulário. A grande vantagem do sistema KWIC está na velocidade do processamento de entrada dos documentos. Assim sendo, é o instrumento que hoje consegue se manter mais próximo do rápido desenvolvimento dos campos científicos e técnicos. Um grande número de serviços de informação publicam bibliografias indexadas pelo sistema KWIC: Bioresearch Index, Biological Abstracts, Chemical Abstracts, Chemical Titles, etc.

Com o aperfeiçoamento e o aumento da capacidade dos equipamentos de processamento de dados, o tipo de lógica empregado no sistema KWIC mostrou estar bem aquém da real capacidade dos computadores. Desenvolveram-se então sistemas usando um maior grau de lógica, baseando-se:

– na estatísticas das palavras do texto;
– na determinação de pesos para termos, de acordo com sua importânica no assunto;
– na freqüência com que determinadas palavras ocorrem juntas nas frases ou parágrafos.

Estas pesquisas foram desenvolvidas na década de 60, mas como alguns problemas não conseguiram ser superados, na década de 70 foram poucos os pesquisadores que se dedicaram ao assunto.

Outra corrente que tem estudado a possibilidade de aplicação dos computadores à análise de texto é baseada na teoria da gramática transformacional de CHOMSKY e HARRIS, citados por FISHMAN (8). O modelo transformacional acredita que existe um conjunto finito de sentenças-padrão *(Kernel Sentences)* para as quais todo conjunto infinito de sentenças em linguagem natural poderia ser transformado. Para CHOMSKY, todas as línguas têm uma estrutura superficial e uma estrutura profunda. Esta última é relativa ao modo como as idéias são formadas na mente humana e portanto seria universal, enquanto que a estrutura superficial poderá variar de acordo com as línguas. O problema então é a conversão da estrutura superficial em estrutura profunda através do procedimento da gramática transformacional. Ao colocarmos todo texto nos termos de sua estrutura profunda estaremos realmente analisando este texto, chegando às sentenças-padrão do mesmo. Poderíamos programar então os computadores para aplicar o modelo transformacional aos textos e teríamos análise automática dos mesmos. Realmente as experiências

realizadas por HARRIS na Universidade de Pennsylvania chegaram a conjuntos de sentenças-padrão de documentos, que muito se assemelhavam aos resumos dos textos feitos com intenção de disseminação de informações.

No entanto, todos esses estudos são experimentais: a possibilidade da aplicação efetiva de computadores à análise de textos ainda está ligada à necessidade de aperfeiçoamento desta área da linguística.

Subject analysis as a basic operation for information retrieval. Different levels of subject analysis: a) to determinete the subject content of a document; b) to interpretate the user's information needs; c) to select or create the information language more adequate to the system. Recomendations to be observed an each situation. Automatic subject analysis.

5. BIBLIOGRAFIA

(1) UNESCO. *Unisist: Indexing principles.* Paris, 1975.

(2) FOSKETT, A.C. *Abordagem temática da informação*. São Paulo, Polígono, 1973, 347p.

(3) MEADOW, C.T. *The analysis of information systens.* 2 ed. Los Angeles, Melville, 1973. 420p.

(4) HATT, Frank. *The reading process.* London, Dive Gingley, 1976. 124p.

(5) LANCASTER, F.W. *Vocabulary control for information retrieval.* Washington, Information Resources Press, 1972. 233p.

(6) GOODMAN, F. *The role and function of the thesaurus in education.* In: THESAURUS of Eric descriptors. 2 ed. New York, C.C.M., 1970.

(7) LUHN, H. P. Keyword in contex for technical literature (KWIC INDEX) *Am. Doc. 9* (4): 288-95, oct. 1960.

(8) FISHAMN, Marilyn, The transformational model of language and information retrieval. *Drexel Library Journal, 8* (2): 193-200, apr. 1972.