Artigo | *Article*

Ciência da Informação em Revista
ISSN: 2358-0763

Indexação automática de documentos utilizando termos extraídos de grifos

Automatic document indexing using terms extracted from highlights

Nathaly Leite
Mestra em Ciência da Informação
Universidade de Brasília, Brasil
nathalyrocha@ibict.br

Dalton Lopes Martins
Doutor em Ciência da Informação
Universidade Federal Fluminense, Brasil
dmartins@gmail.com

Resumo

O processo de indexação e recuperação de informações desempenha um papel crucial na organização eficiente de documentos e no acesso rápido a conteúdo relevante. A pesquisa aborda a automatização no processo de indexação de documentos, abordando também aspectos sobre extração de termos a partir de grifos. A metodologia de trabalho possui abordagem qualitativa, enquadrando-se no tipo de revisão de literatura caracterizada como narrativa, visto que foi realizada de forma indutiva, a partir da busca de levantamento bibliográfico com busca de termos relacionados às temáticas do trabalho em bases de indexação. Os resultados enfatizam a importância da pesquisa contínua e do desenvolvimento de algoritmos eficazes para a indexação automática de documentos com base em termos extraídos de grifos. Essa abordagem tem o potencial de aprimorar a eficiência da organização da informação e tornar o acesso a documentos mais preciso e ágil em diversas áreas, incluindo Biblioteconomia, Ciência da Informação e sistemas de gerenciamento de documentos.

Palavras-chave: indexação; automatização; destaques em textos.

Abstract

The process of indexing and retrieving information plays a crucial role in the efficient organization of documents and quick access to relevant content. The research addresses automation in the document indexing process, also discussing aspects related to extracting terms from highlights. The working methodology takes a qualitative approach, fitting into the category of literature review characterized as narrative, as it was conducted inductively through a bibliographic search for terms related to the themes of the work in indexing databases. The results emphasize the importance of continuous research and the development of effective algorithms for the automatic indexing of documents based on terms extracted from highlights. This approach has the potential to enhance the efficiency of information organization and make document access more accurate and agile in various fields, including Library Science, Information Science, and document management systems.

Keywords: *indexing; automation; text highlights.*

doi: 10.28998/cirev.2024v11e16800

Submetido em: 07/11/2023
Aceito em: 30/11/2024
Publicado em: 25/12/2024

1 INTRODUÇÃO

A indexação automática, um elemento-chave no universo da recuperação de informações, tem vivenciado uma notável evolução ao longo das últimas décadas. Em um mundo onde a quantidade de dados digitais continua a crescer a taxas exponenciais, a necessidade de sistemas eficazes de organização e busca de informações se tornou mais urgente do que nunca. A indexação automática, que engloba a atribuição de metadados e a categorização de conteúdo de maneira automatizada, desempenha um papel fundamental na otimização da recuperação de informações em domínios que vão desde motores de busca na web até sistemas de gerenciamento de documentos e bibliotecas digitais.

À medida que a tecnologia evolui, surgem novas oportunidades e desafios para a indexação automática. Neste artigo de revisão de literatura, exploraremos a trajetória dessa evolução, desde suas origens históricas até as técnicas de ponta que estão moldando o cenário atual. Através da análise de pesquisas e avanços na área, apresentaremos uma visão abrangente das principais tendências e conquistas que têm impulsionado a eficiência e a eficácia da indexação automática.

Um aspecto notável deste artigo é a inclusão de trechos grifados por leitores, que realçam os pontos de interesse e as percepções de especialistas e estudiosos na área. Essas marcações destacam não apenas os aspectos mais relevantes do texto, mas também oferecem *insights* valiosos sobre as perspectivas e opiniões de quem se dedica ao estudo da indexação automática. Ao incorporar esses destaques, esperamos enriquecer ainda mais a compreensão desse campo em constante evolução e fornecer aos leitores uma visão abrangente do estado atual.

2 REFERENCIAL TEÓRICO

A relação da sociedade com a produção e consumo de informações vem se alterando ao longo do tempo. Foi convencionado chamar a configuração dessa relação atualmente de sociedade da informação, a qual segundo Gouveia e Gaio (2004) é uma sociedade que predominantemente utiliza as tecnologias de informação e comunicação para a troca de informações em formato digital e que suporta a interação entre indivíduos e organizações com recurso a práticas e métodos em construção permanente. A internet ganha papel de destaque nas relações informacionais e nos processos de armazenamento, processamento e comunicação da informação, visto que “indivíduos e organizações podem produzir e consumir informação de um modo quase instantâneo e a qualquer hora e em qualquer lugar” (Silva; Gouveia, 2020, p. 16).

A internet e a *web* mundial têm fundamental importância nessa sociedade da informação, já que segundo Silva (2006), permitem a conexão de bilhões de pessoas em todo o mundo por meio de dispositivos tecnológicos. A pandemia da Covid-19 que aconteceu em 2020 intensificou o uso da internet como ferramenta para estudo, trabalho, registro e utilização de informações. Atualmente, o mundo conta com 5,3 bilhões de usuários da rede mundial de computadores, segundo pesquisa publicada pela ONU (2022)[1].

[1] Disponível em: https://news.un.org/pt/story/2022/09/1801381#:~:text=Ao%20todo%2C%20existem%205%2C3,da%20pandemia%20de%20Covid%2D19.

Além da rede mundial, também os *gadgets* de acesso à informação mudaram o paradigma da leitura em papel para o digital. No Brasil, a Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros - TIC Domicílios (2020) mostra que o acesso à tablets com conexão à internet compreende 8% da população, e os que acessam por notebook representam 30%. Outra pesquisa, esta do *Pew Research Center* (2022), mostra que apesar dos livros físicos continuarem sendo a preferência da maioria dos leitores adultos, o percentual daqueles que leem em formato digital já é de 30%.

No âmbito da literatura científica também é perceptível o crescimento da quantidade de documentos disponíveis em formato digital. A comunidade acadêmica pode acessar milhares de publicações, livros e periódicos indexados em bases de dados diversas. Para que esse acesso aconteça, destaca-se o trabalho de indexação, apontado por Gil Leiva (2009) como primordial para a recuperação da informação científica. Com a crescente produção de conhecimento atual, torna-se necessário o uso de ferramentas tecnológicas que proporcionem mais agilidade aos processos de documentação. Se tratando da indexação, Lancaster (2004) e Moreiro González (2004) apontam que as automatizações das tarefas envolvidas corroboram para um importante ganho de tempo dos profissionais indexadores que atuam em unidades de informação.

A influência do meio digital é percebida não só pelo acesso rápido e facilitado aos documentos em bases de dados, como também é notória a mudança na forma de interação dos usuários da informação com os textos em novo formato. A leitura de documentos digitais se dá através de telas de computador ou outros dispositivos como *tablets*, celulares *smartphones* e *e-readers*, que são aparelhos eletrônicos desenvolvidos especificamente para leitura de documentos em formato digital. Estes dispositivos apresentam funções de interação que antes não eram experimentadas em mídias analógicas, a exemplo do uso de *links* e *hiperlinks*, o que permite um enriquecimento informacional na medida em que complementa o texto com outros textos que estão acessíveis ao leitor com um clique na tela do dispositivo utilizado.

Outra forma de interação do leitor com os documentos digitais é o manuseio da informação sem que se prejudique a integridade do texto: em um documento lido em uma tela, é possível riscar, grifar, anotar, marcar páginas sem que se perca o material original utilizado. Essa possibilidade de interação pode ser investigada por diversos aspectos, no presente estudo considera-se um aspecto relevante para a indexação, mencionado por Glushko (2016). Ao elencar elementos que podem ser considerados em um processo de organização da informação, o autor cita os destaques feitos por um leitor ao grifar ou destacar[2] um texto como um recurso organizacional, chamado pelo autor de recurso de interação.

3 METODOLOGIA

Para a construção deste trabalho foi utilizada abordagem qualitativa, enquadrando-se no tipo de revisão de literatura caracterizada, segundo Sampieri, Collado e Lucio (2013), como revisão narrativa, visto que foi realizada de forma indutiva, a partir da busca de levantamento bibliográfico com busca de termos relacionados às temáticas do trabalho em

[2] No texto original o autor utiliza o termo *highlight*, em tradução nossa entendemos equivalência semântica em português com os termos grifar ou destacar.

bases de indexação. Esta abordagem de revisão se caracteriza pela fluidez ao descrever, analisar e sintetizar o conhecimento existente sobre um determinado tema. Ao contrário de revisões sistemáticas que seguem protocolos rígidos, a revisão narrativa permite uma abordagem mais flexível, enfatizando a interpretação e contextualização dos estudos revisados. Suas características incluem uma seleção ampla de fontes, uma organização mais livre do material, foco na compreensão dos aspectos qualitativos e interpretativos dos estudos. Os textos utilizados foram recuperados principalmente através da busca na base de periódicos Capes, Brapci, Scielo e Web of Science, além da busca direta por trabalhos citados nas leituras realizadas.

4 RESULTADOS

4.1 Documentos digitais

A autora Blanca Rodríguez Bravo em seu livro *El documento: entre la tradición y la renovación* (2002) defende que o documento digital continua sendo documento, com a especificidade de que a união da mensagem com o suporte não é inseparável, como acontece no documento analógico.

Contudo, é necessário que se estabeleça as características específicas do documento produzido e utilizado em meio digital, já que as características de um suporte impactam nos processos que garantem a circulação, recuperação, utilização e preservação de um documento. Se é preciso tratar algo, é necessário entender o que é algo. Os esforços empenhados por Paul Otlet e Henri La Fontaine para definir o que é um documento estão intimamente ligados ao desejo de documentar os objetos informativos com os quais lidavam, com objetivo de possibilitar a recuperação da informação por quem desejasse fazê-la. Nesse sentido, visto que os meios pelos quais as informações são recuperadas atualmente, é imprescindível que se entenda também como documentar em meio digital.

Para Buckland (1997), a mudança para o meio digital representa uma virada que possibilita a redefinição do próprio documento. Além de alargar o conceito de documento para incluir elementos não textuais, a exemplo de imagens, a ideia de redefinição proposta por Buckland seria depois endossada por López Yepes (1998), que aponta que o documento digital é um objeto informativo livre das limitações de suportes tradicionais, como o papel, já que a circulação em meio eletrônico é facilitada. Para o autor, o documento digital consolida de vez a ideia de documento.

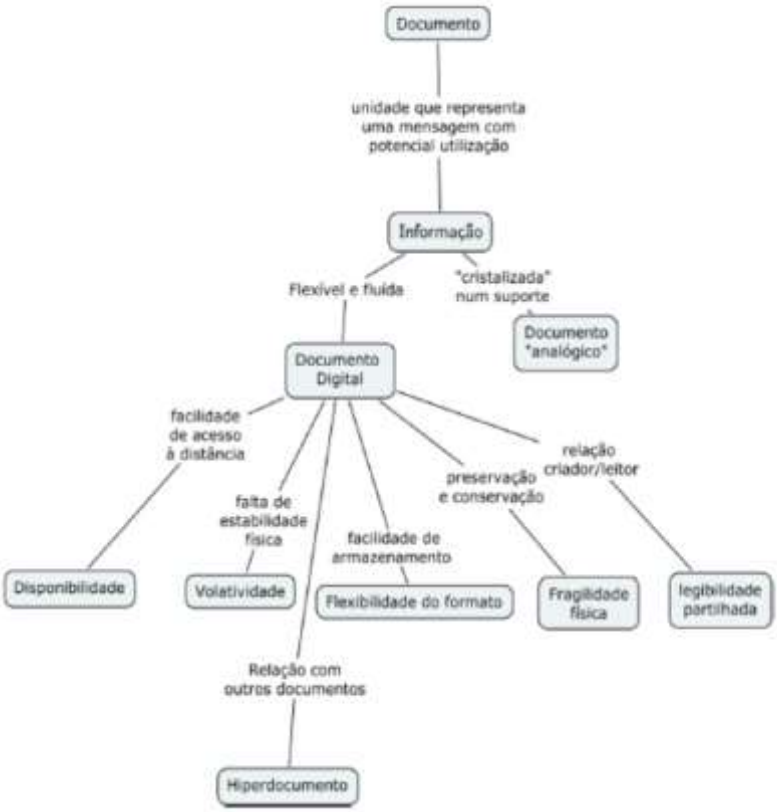
Outro autor que escreveu sobre a temática e trouxe características do documento digital foi Michel (2000), o qual aponta características como: a facilidade de armazenamento em comparação com os documentos analógicos, o que possibilita facilidade também nos processos de localização e recuperação; a disponibilidade de acesso à distância de forma instantânea e a flexibilidade de formato. Outros pontos destacados pelo autor são a possibilidade de relacionar documentos digitais; a versatilidade para registrar informações e se adequar às necessidades da comunicação humana dos tempos atuais e, por fim, a necessidade dos profissionais que lidam com informação adequarem os seus trabalhos aos meios tecnológicos.

Santos e Flores (2012), em uma proposta de definição, dizem que “[...]um documento digital é a informação registrada em suportes acessíveis por meio de um equipamento computacional.” Os autores apontam benefícios encontrados nos documentos digitais, tais

como: redução de custos e a otimização da criação, tramitação e difusão dos documentos; novas possibilidades de interação entre pessoas, documentos e instituições de informação e a adaptação dos novos usuários frente ao mundo digital, com familiarização que facilita o acesso e uso dos documentos e suas informações. Com isso, os autores também citam a coexistência de documentos em suportes analógicos e digitais, excluindo a substituição do primeiro pelo segundo.

Siqueira (2012) realizou trabalho de revisão de literatura com o objetivo de levantar os autores e conceitos que abordam o documento digital, a fim de sistematizar as ideias que se têm sobre o assunto. A autora, ao fim do trabalho de revisão, apresenta um diagrama de conceitos elaborado em forma de mapa mental, interligando as definições que encontrou durante a revisão feita. O diagrama está replicado na imagem abaixo:

Figura 1 - Mapa conceitual com definições de documentos



Fonte: Siqueira, Jessica Campos (2012).

A mudança de paradigma dos documentos passando do analógico para o digital também implica em mudanças na relação do leitor com tais documentos. A próxima subseção é dedicada a elucidar importantes aspectos da leitura em meio digital.

4.2 Leitura em meio digital

Na medida em que as tecnologias de informação e comunicação foram incorporadas no dia a dia das pessoas, e considerando também a criação de documentos, imagens, filmes e outros tipos de documentos em formatos digitais, a forma de consumo dos usuários de informação também se adequou a este contexto. Um estudo cientométrico realizado por Dantas *et al.* (2017) sobre o

cenário das pesquisas sobre leitura mostra o aumento de publicações sobre o tema em 2009, o que é explicado pela adoção de novos meios de leituras tais como computadores, *tablets* e *smartphones*. Os autores também notaram que o aumento de pesquisas sobre leitura coincide com o lançamento de dois *gadgets* de leitura digital populares: o *Kindle*, leitor eletrônico da empresa *Amazon*, e o *iPad*, *tablet* da empresa *Apple*.

O aumento da população leitora em meios digitais continua crescente. Um relatório produzido pela *BusinessWire*[3] aponta tendência de aumento em 28% das receitas geradas pelo mercado global de *e-books*, acrónimo para *eletronic book*, os livros digitais. No Brasil, uma pesquisa realizada pela Câmara Brasileira do Livro com base no ano de 2021, publicada em 2022[4], corrobora a tendência global, apresentando um crescimento geral de 23% no faturamento das vendas de livros digitais, o que representa 6% do mercado editorial no país.

A leitura digital requer do leitor diferentes competências daquelas que a leitura em documentos analógicos demanda. Agora, além do processo básico de interpretação e compreensão de signos, também se mostra necessária a adaptação ao dispositivo digital utilizado para a leitura, o que é chamado por Zayas (2010) de alfabetização digital. Outro fator diferencial da leitura em meio digital, mencionado na seção anterior, é o hipertexto. Além de alterar a estrutura do documento, a leitura digital altera também a forma que o leitor lida com o conteúdo, que agora pode fazer vários itinerários e rotas para a leitura, não mais seguindo o fluxo linear imposto em mídias analógicas.

É natural também observar mudanças na interação do leitor com o livro quando se muda o suporte da leitura. Os aparelhos utilizados para tais leituras desempenham um papel fundamental na determinação de como um leitor vai interagir com um texto, na medida em que cada dispositivo vai disponibilizar diferentes funções de interação: cliques em *links* que são acessados na internet, possibilidade de grifar, anotar ou destacar, salvar comentários, por exemplo. Estas funções, inclusive, são apontadas como benefícios da leitura digital em detrimento da leitura em meio físico, segundo Kelly (2011), o que o livro sempre quis foi ser anotado, marcado, sublinhado, ter as pontas de suas páginas dobradas, ser resumido, ganhar referências cruzadas, hiperlinks, ser compartilhado, e dialogar. Ser digital lhes permite fazer tudo isso e muito mais.

A UNESCO (2013) também apresenta um conceito relacionado à leitura em meio digital denominado em inglês *mobile learning* para se referir ao conjunto de práticas de ensino possibilitadas pela utilização da tecnologia digital móvel, combinada ou não com outras tecnologias de informação e comunicação, e entre as práticas se encontra a leitura digital. Um estudo de Bernardo e Karwoski (2017), realizado com estudantes de graduação do curso de Letras, mostrou que 50% dos participantes escolheram, durante o período do experimento, ler textos acadêmicos em dispositivos digitais. Os autores perceberam o hábito de praticar a leitura de textos para aprendizado utilizando o meio digital entre os alunos:

> Quanto a textos menores, como artigos acadêmicos, não titubeamos em afirmar que não só a preferência como já hábito formado é a leitura em arquivo digital via telefone celular. Os professores mesmos, segundo os participantes, estimulam essa

[3] Disponível em: https://www.researchandmarkets.com/reports/4894553/global-digital-publishing-market-2022-2026

[4] Disponivel em: https://cbl.org.br/pesquisas_de_mercado_categoria/1-producao-e-vendas-do-setor-editorial-brasileiro/

> forma de leitura a partir da disponibilização de textos e materiais didáticos *on-line*. (Bernardo; Karwoski, 2017, p. 804).

Para além do meio utilizado para realizar a leitura, existe também o conjunto de comportamentos executados por um leitor durante a leitura, conjunto este chamado de estratégias de leitura (COOK; MAYER, 1983). Além da influência do meio escolhido para leitura, seja digital ou analógico, o que define o comportamento adotado durante uma leitura é o objetivo final do leitor. Em uma leitura com objetivo de aprendizado, as práticas mais comuns são: sublinhar palavras-chave, grifar passagens do texto, anotar e sumarizar tópicos. A próxima subseção destrincha um destes comportamentos adotados por leitores: as anotações.

4.3 Anotações: conceitos e aplicações

Anotar é uma estratégia de leitura para auxiliar o leitor na compreensão do texto. DiVesta e Gray (1972) apontam duas funções distintas possíveis de serem executadas pelo ato de anotar durante uma leitura e/ou aula: armazenamento e codificação. A primeira função caracteriza as anotações como um armazenamento externo de informações, que podem servir como meio de revisão posterior à leitura inicial. Já a função de codificação sugere que o processo de anotar estimula o aumento da concentração, da capacidade de análise sobre a leitura, do desenvolvimento de ideias próprias sobre o assunto tratado e da organização de informações.

A anotação possui também um sentido de enriquecimento informacional de documentos, como proposto no conceito de anotação semântica, cujo objetivo é possibilitar que um documento digital criado para interpretação humana possa também ser interpretado por máquinas. Através da adição de palavras com significados relacionados à temática do texto, as anotações incluídas no documento permitem a recuperação da informação em sistemas de buscas precisos. (Fontes *et al*., 2010).

4.4 Destaques em texto

Neste trabalho, utilizaremos os termos destaque ou destacar como sinônimos para grifo ou grifar, respectivamente. O termo destacar, uma tradução do termo em inglês *highlight*, tem significado definido pelo dicionário Merriam Webster como *to center attention on: emphasize, stress*; em português significa salientar algo que merece atenção. Já o termo grifar, definido em português pelo dicionário Michaelis como escrever com grifo; sublinhar, nos aproxima da ideia de destacar termos em um texto, justamente o contexto definido no estudo proposto.

Assim como as anotações, destacar informações em um texto através de grifos também está intimamente ligado à compreensão da leitura, segundo Leutner et al. (2007). Os autores apresentam a categorização de estratégias de aprendizado por leitura em dois níveis: 1) superficial; e, 2) profundo. No nível superficial, o leitor se concentra em memorizar as informações apenas lendo o texto, enquanto no nível profundo o leitor seleciona e estrutura as informações mais importantes, adicionando anotações para relacionar o conteúdo com outros já conhecidos e criando assim um esquema mental para a absorção do novo conteúdo lido.

O processo de selecionar e destacar as informações que o leitor considera importantes está ligado ao segundo nível de aprendizado, o mais profundo. Ainda segundo Leutner et al. (2007), o ato de grifar auxilia o processo de armazenar e recuperar informações, num contexto mental do leitor. Weinstein e Mayer (1986) também escrevem sobre a técnica de destacar informações e apontam duas funções para a prática: 1) identificar e focar a atenção do leitor nas informações importantes de um texto; e 2) armazenar as informações importantes identificadas na memória de curto prazo, para que seja processada posteriormente.

4.5 Organização da informação

A informação, enquanto objeto de estudo da CI, é um termo bastante discutido com múltiplas definições. Uma miríade de estudos aborda seus diversos aspectos e possíveis formas de análise. Para citar um exemplo das definições levantadas em tais estudos, Le Codiac (2004) diz que a informação é um conhecimento registrado que comporta um elemento de sentido, registro este que é feito graças a um sistema de signos, a linguagem. Com base nas definições e abordagens existentes, emergem formas de exploração e estudo da informação, dentre as quais está a organização da informação (OI). Aguiar e Kobashi (2013), sobre a organização, afirmam que "no domínio da CI, ela pode ser compreendida como uma série de atividades processuais com a finalidade de descrever intelectualmente conteúdos documentais para serem representados nos sistemas de recuperação da informação".

Organizar informações já era uma preocupação muito antes da consolidação da CI como área de estudo, visto que Otlet (1934) já considerava a prática de organizar dentro dos processos da documentação. Para o autor, o objetivo principal era: "A organização da documentação em uma base cada vez mais abrangente, cada vez mais prática, de forma a alcançar para o trabalhador intelectual o ideal de um mecanismo para explorar o tempo e o espaço".

Conceituando o termo organização da informação, Café e Sales (2010) definem que é um processo de arranjo de acervos tradicionais ou eletrônicos realizado por meio da descrição física e de conteúdo (assunto) de seus objetos informacionais.

É possível mencionar diversas técnicas desenvolvidas neste âmbito organizacional para que se alcance os objetivos mencionados acima, como a classificação e a indexação, por exemplo. A classificação permite representar a informação em forma de símbolos de classificação e números, descrevendo o conteúdo de forma abrangente. Já a indexação busca descrever o conteúdo de forma específica, gerando termos e palavras-chave que possam atender possíveis perguntas de usuários em suas buscas por informação. As duas práticas atuam de forma complementar.

A partir de práticas como estas citadas, torna-se possível o desenvolvimento de sistemas e ferramentas organizadas para o uso no processo de recuperação da informação. Para mencionar alguns dos importantes avanços para o campo da organização da informação, temos a criação do sistema de Classificação Decimal de Dewey, em 1876, que propõe a classificação por assuntos de maneira hierárquica. Outro sistema de classificação de grande importância nesse campo é a Classificação Decimal Universal, proposta por Paul Otlet e Henri La Fontaine em 1905, que possibilita a organização temática por meio de códigos numéricos.

Com as contribuições dos estudos do bibliotecário Ranganathan, um novo método surgiu no campo da organização informacional. O estudioso foi responsável pela formulação, entre 1933 e 1960, do Colon Classification, um método que extrapola as relações hierárquicas apresentadas em outros sistemas de classificação existentes até o momento e permite a categorização de assuntos que se agrupam por características similares, possibilitando a criação de relações entre estes. Além do citado método, foi Ranganathan quem proporcionou arcabouço teórico para posteriores construções de códigos de classificação diversos ao descrever seu modelo de princípios de categorias fundamentais, o PMEST (acrônimo de *Personality, Matter, Energy, Space e Time*).

A influência dos estudos dos autores mencionados nesta seção, considerados pioneiros no contexto da organização da informação, perduram nas práticas e estudos atuais. A exemplo, podemos citar os termos trazidos à luz por Ranganathan em seu modelo teórico que ainda hoje são utilizados como formas de classificação em diversos sistemas de recuperação da informação, tais como facetas, níveis e focos. Os resultados possibilitados e alcançados por tais estudos vão ao encontro da construção do conhecimento, que é construído a partir da circulação e uso da informação, sua matéria-prima.

A organização da informação está intimamente ligada ao processo de descrever objetos informacionais. Shera e Egan (1953) apontam que a descrição é o processo de "individualização de determinado item entre o vasto número dos que formam o conjunto de literatura". Essa descrição se dá em dois níveis: o conceitual e o físico. No nível conceitual, é necessário descrever o conteúdo do objeto informacional, o conhecimento registrado. Já no nível físico, a descrição se encarrega de especificar as características do suporte no qual o conhecimento está registrado. Destes dois tipos de descrições, derivam-se alguns processos e práticas de organização da informação.

Baseado na sistematização dos elementos gerais da organização da informação realizada por Brascher e Monteiro (2010), o esquema abaixo, seguido de breve definição dos termos elencados, foi elaborado com o intuito de ilustrar os procedimentos realizados no âmbito da OI e suas ligações com os níveis de descrição supramencionados.

A revisão realizada focou nos aspectos de um dos processos acima representados: a indexação, definido por Lapa e Corrêa (2014) do seguinte modo:

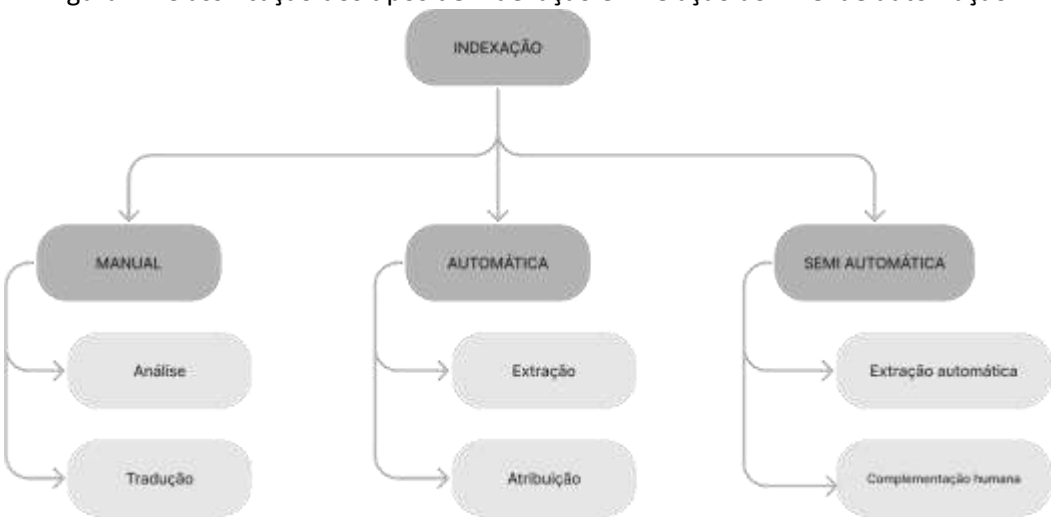
> A indexação é um processo de tratamento temático essencial, pois consiste no ato de identificar e descrever um documento de acordo com o seu assunto, e cujo principal objetivo é orientar o usuário sobre esse conteúdo intelectual, permitindo, dessa forma, a sua recuperação de forma ágil e eficiente. (LAPA E CORRÊA, 2014)

4.6 Indexação

A indexação, segundo Guinchat e Menou (1994), é a operação pela qual se escolhe os termos mais apropriados para descrever o conteúdo de um documento. A prática de atribuir termos que identifiquem um documento a partir das informações as quais ele carrega é de suma importância para a posterior recuperação da informação através de sistemas. A escolha de termos deve ser realizada de forma que atenda aos interesses do usuário que utilizará os sistemas de recuperação da informação. Assim, um mesmo documento pode ser indexado a partir de termos diferentes, a depender de quem realizará consultas (Lancaster, 2004). A seguir, na Figura 2, apresenta-se uma classificação dos tipos de indexação, considerando o tipo de indexação em relação à tecnologia utilizada e as etapas envolvidas.

No processo de indexação manual, Lancaster (2004) aponta duas etapas de trabalho: a análise conceitual e a tradução. Na etapa de análise, o profissional indexador realiza a leitura do documento a fim de compreender seu conteúdo e definir a temática tratada, definir o assunto do documento. É nessa etapa onde o indexador deve tomar decisões conceituais sobre como sintetizar o conteúdo da melhor forma para atender as necessidades informacionais específicas do grupo de usuários que será atendido. Aqui, o autor menciona que é importante a formulação de perguntas por parte do indexador tais como: de que se trata o documento? Por que foi incorporado a este acervo? Quais de seus aspectos são de interesse dos usuários?

Figura 2 - Classificação dos tipos de indexação em relação ao nível de automação



Fonte: adaptado de Lancaster (2004) e Pinto (2001);

A indexação orientada ao usuário, termo esse cunhado por Fidel (1994), abarca diversos pontos de questionamentos acerca de como atender satisfatoriamente às necessidades informacionais dos variados grupos de usuários que possam existir, e que são citados também na obra de Lancaster (2004). O primeiro ponto levantado é sobre a competência de conhecimento do público que fará uso do conjunto de documentos indexados, como um importante complemento do conhecimento das técnicas de indexação, indispensáveis ao trabalho. Outro ponto elucidado por Lancaster é a impermanência da indexação realizada, visto que os interesses de um usuário ou um grupo de usuários pode e provavelmente mudará ao longo do tempo. Como exemplo, aponta-se um grupo de pesquisadores que avançam em seus estudos e precisam organizar seus documentos com frequência para que continuem tendo suas necessidades informacionais atendidas.

Já na etapa seguinte, a de tradução, a tarefa do indexador consiste em sintetizar o assunto do documento, identificado na etapa de análise, em termos indexáveis que serão incluídos na base de dados. É nesta etapa que Lancaster (2004) apresenta os conceitos de indexação por atribuição e por extração. Na indexação por extração, o indexador retira do próprio texto a ser indexado alguns termos (palavras ou expressões) que considere mais representativos para realizar a indexação. Já na indexação por atribuição o indexador seleciona termos que não estão presentes no texto, mas que podem ser considerados relevantes dada a temática ou o interesse dos usuários. Nesse segundo tipo pode se utilizar

ferramentas como o vocabulário controlado, que é uma lista de termos autorizados que se convencionou usar em um determinado contexto, como em uma instituição, por exemplo.

A indexação automática, realizada com apoio de ferramentas computacionais, é estudada desde a década de 1950, através da análise automatizada de textos. Os estudos, conforme Robredo (2005), mostram que o computador é uma ferramenta indispensável para garantir acesso rápido à bases de dados científicos, auxiliando nas tarefas de processamento de dados e informações. Lapa e Corrêa (2014) conceituam a indexação automática como um conjunto de operações matemáticas, linguísticas e de programação que, quando aplicada na análise de documentos, faz o processamento dos conteúdos, selecionando automaticamente os termos representativos dos assuntos destes documentos para serem utilizados posteriormente em processos de recuperação de informação.

São identificados dois tipos de indexação automática: por atribuição e por extração, ambos referentes ao modo de extração dos termos de um documento textual. Lancaster (2004) define a indexação por atribuição como o processo de atribuir automaticamente termos oriundos de um vocabulário controlado de acordo com os conjuntos de palavras e expressões encontrados após a análise textual do documento. Neste tipo de indexação o vocabulário controlado possui influência no processo de análise dos documentos, ou seja, a escolha dos termos para atribuição está intimamente ligada e condicionada ao vocabulário que será utilizado. Segundo Lancaster, este é o tipo mais difícil de se realizar de forma computacional pois é necessário "[...] desenvolver, para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer frequentemente nos documentos [...]" (Lancaster, 2004).

Na abordagem de indexação por extração, os termos selecionados para indexar um documento são escolhidos dentro do próprio texto, usando a linguagem natural ao invés de termos estruturados em um vocabulário. Aqui, considerando os documentos digitais, é possível utilizar *softwares* computacionais para realizar um processo parecido com o desenvolvido por um indexador humano. Os termos extraídos podem atender aos critérios de frequência ou posição das palavras dentro do texto ou no resumo além da possibilidade de extração de termos considerando também o contexto do documento (Lancaster, 2004). Aqui, ao contrário da indexação por atribuição, o autor explicita o ótimo desempenho computacional na tarefa de extração, considerando-a bastante coerente quando validada.

Ainda no âmbito da indexação realizada com auxílio de computadores, além dos dois tipos de indexação já mencionados, existe também a indexação semiautomática. Segundo Pinto (2001), esta é um tipo de indexação que une etapas da indexação manual e da indexação automática, de forma que a indexação através da extração de termos realizada por um *software* seja revisada e validada por um indexador humano. O *software* SISA mencionado na figura 4 é uma ferramenta para realização deste tipo de indexação com uma relacionada ao processo em que o sistema computacional realiza a atividade de análise do conteúdo do documento e, posteriormente, um indexador humano avalia os termos para indexação propostos pelo sistema (Narukawa, 2011).

4.7 Indexação automática e destaques em textos

A revisão realizada apontou alguns indícios de relação entre os grifos e as práticas de organização da informação, sendo que de modo geral, a etapa de leitura e marcação de um texto pode ser enxergada como uma fase preliminar das atividades de organização da

informação, em específico, daquelas voltadas à descrição conceitual dos documentos (Brascher; Monteiro, 2010). Podemos destacar a ligação mais direta entre alguns:

- Indexação: a marcação de conceitos, termos técnicos e pensamentos do autor são de grande valia no processo de indexação, visto que funcionam como uma seleção preliminar de trechos que caracterizam o documento lido e que podem funcionar como um refinamento para destacá-lo de forma individual dentro do corpus em que se encontra;
- Resumo: para a elaboração de resumos, as marcações em partes estruturais do texto como os objetivos e resultados é de extrema utilidade para promover agilidade, visto que a identificação de tais partes já estarão destacadas;
- Classificação: as marcações de trechos que indicam ligações de um documento com outros de mesma temática podem ser apoio para a classificação do documento dentro de um sistema com documentos diversos.

Também cabe destacar certa sinergia com o que trazem as autoras Brascher e Café (2008) ao formularem que a OI se preocupa em entender de forma individual a estrutura dos objetos informacionais para organizá-los sistematicamente. Ao destacar as partes principais de um texto, ou quando destaca estruturas importantes como objetivos e resultados, o leitor está destrinchando o objeto informacional e proporcionando um possível caminho para a organização de uma coleção, ou do próprio objeto dentro de uma coleção já estruturada.

5 CONCLUSÃO

A indexação automática de termos grifados em textos é uma área de pesquisa fundamental que desempenha um papel crucial na organização e recuperação eficiente de informações em ambientes digitais. Ao longo deste artigo de revisão, examinamos as principais técnicas e abordagens utilizadas para identificar e indexar automaticamente termos grifados em textos. Compreendemos que, embora existam diversos desafios a serem superados, como o reconhecimento de entidades e a interpretação de contextos, as soluções estão em constante evolução.

A indexação automática de termos grifados oferece inúmeras vantagens, como a melhoria da navegabilidade em documentos extensos, aprimoramento da pesquisa e recuperação de informações, além de facilitar a categorização e organização de conteúdos digitais. No entanto, é importante ressaltar que a qualidade e a precisão da indexação dependem da escolha adequada das técnicas e do treinamento de algoritmos, bem como da manutenção regular das bases de dados.

À medida que a tecnologia continua a avançar, é provável que a indexação automática de termos grifados em textos se torne ainda mais eficaz, beneficiando a academia, as empresas e os usuários em geral. Portanto, esta área de pesquisa permanece promissora e continuará a desempenhar um papel importante na gestão e exploração de informações em um mundo cada vez mais digital.

REFERÊNCIAS

AGUIAR, F.; KOBASHI, N. Organização e representação do conhecimento: perspectivas de interlocução interdisciplinar entre Ciência da Informação e Arquivologia. In ENCONTRO

NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16.,2013, Florianópolis. **Anais** [...] Florianópolis: PPGCI, 2013

BERNARDO, J. C. O.; KARWOSKI, A. M. **A leitura em dispositivos digitais móveis.** ETD - Educação Temática Digital, [S. l.], v. 19, n. 4, p. 795–807, 2017. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/8646355.

BRASCHER, Marisa; CAFÉ, Ligia. Organização da informação ou organização do conhecimento? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais** [...] São Paulo: PPGCI, 2008. p. 1-14

BRÄSCHER, M.; MONTEIRO, F. S. Organização da informação em repositórios digitais. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 15, n. 29, 2010.

BRAVO, B. R. **El documento: entre la tradición y la renovación.** Ediciones Trea, 2002

BRITO SILVA, S.R. **Sistemas de indexação automática por atribuição:** uma análise comparativa. Dissertação (Mestrado em Ciência da Informação – Departamento de Ciência da Informação, Universidade Federal de Pernambuco. Recife. 2020.

BUCKLAND, M. K. What is a document? **Journal of the American Society for Information Science**, Washington, v.48, n.9, p. 804-809, Sept., 1997.

CAFÉ, Lígia; SALES, R. **Organização da informação:** conceitos básicos e breve fundamentação teórica. In: Jaime Robredo; Marisa Bräscher (Orgs.). Passeios no Bosque da Informação: Estudos sobre Representação e Organização da Informação e do Conhecimento. Brasília DF: IBICT, 2010. 335 p. ISBN: 978-85-7013-072-3. Capítulo 6, p. 115-129. Edição eletrônica.

COOK, L.K.; MAYER, R.E. Reading Strategies Training for Meaningful Learning from Prose. In: PRESSLEY, M.; LEVIN, J. R. (Ed.) **Cognitive Strategy Research**. New York? Springer, 1983. p. 87-131. http://dx.doi.org/10.1007/978-1-4612-5519-2_4

DANTAS, T., MANGAS-VEGA, A., GOMÉZ-DÍAZ, R., CORDÓN-GARCÍA, J. A. Pesquisa em leitura e pesquisa em leitura digital: panorama do atual cenário científico. **Informação & Sociedade:** estudos., João Pessoa, v.27, n.2, p. 117-131, maio/ago. 2017.

DI VESTA, F. J.; GRAY, G. S. Listening and note taking. **Journal of Educational Psychology**, v. 63, n. 1, p. 8–14, 1972.

FAVERIO, M.; PERRIN, A. Three-in-ten Americans now read e-books. **Pew Research Center**, 6 jan. 2022. Disponível em: https://www.pewresearch.org/fact-tank/2022/01/06/three-in-ten-americans-now-read-e-bo oks/. Acesso em: 25 fev. 2024.

FIDEL, R. User-oriented indexing. **Journal of the American Society for Information Science**, v. 45, p. 572-576. 1994.

GIL-LEIVA, I. **Manual de indización:** teoría y práctica. Gijón: Trea, 2009.

GLUSHKO, Robert J. **The Discipline of Organizing:** Professional Edition. 4. ed. O'Reilly Media, 2016.

PASCUAL, C. H. El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural. In: MOREIRO GONZÁLEZ, José Antonio. **Anales de Documentación**. 2005. p. 285-98. Disponível em: https://revistas.um.es/analesdoc/article/view/3501. Acesso em: 25 fev. 2024.

GOUVEIA, L. M. B; GAIO, S. **Sociedade da informação:** balanço e oportunidades. Rio de Janeiro: Universidade Fernando Pessoa, 2004.

GUINCHAT, C.; MENOU, M. **Introdução geral às ciências e técnicas da informação e documentação**. 2.ed. rev. aum. Brasília: Ibict; CNPq, 1994. 540 p.

KELLY, Kevin. O que os Livros se tornarão. In: SILVEIRA, Julio (Org.). **Livrolivre, Novas Possibilidades do Digital para a Escrita, a Leitura e a Publicação.** Rio de Janeiro: Imã Editorial, 2011. p.17-37.

LANCASTER, F.W. **Indexação e Resumos:** teoria e prática. 2. ed. Brasília, DF: Briquet de Lemos, 2004

LAPA, R. C.; CORRÊA, R. F. Indexação automática no âmbito da Ciência da Informação no Brasil. **Informação & Tecnologia**, v. 1, n. 2, p. 59-76, 2014. Disponível em: http://hdl.handle.net/20.500.11959/brapci/41624. Acesso em: 25 fev. 2024.

LE COADIC, Y.F. **A Ciência da Informação**. 2. ed. Brasília, DF: Briquet de Lemos, 2004

LEUTNER, D.; LEOPOLD, C.; DEN ELZEN-RUMP, V. Self-regulated learning with a text-highlighting strategy: A training experiment. **Zeitschrift für Psychologie/Journal of Psychology**, v . 215, n. 3, p. 174–182, 2007.

LÓPEZ YEPES, J. Hombre y documento: del homo sapiens al homo documentador. **Scire**, Zaragoza, v.4, n.2., jul-dec., 1998.

MICHEL, J. L'Information et documentation un domaine d'activité professionnelle en mutation: LCN Les Métiers du Numérique. **Hermès**, v. 1, n.3, p. 47-64, 2000.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação Automatizada de Artigos de Periódicos Científicos: análise da aplicação do *software* SISA com uso da terminologia DeCS na área de Odontologia. **Informação & Sociedade:** estudos, João Pessoa, v. 19, n.2, p. 99-118, 2009.

OTLET, P. **Traité de documentation:** le livre sur le livre: théorie et pratique. Bruxelles: Mundaneum, 1934.

PINTO, V. B. **Indexação documentária: uma forma de representação do conhecimento registrado.** Perspectivas em Ciência da Informação, Belo Horizonte, v.6, n.2, p.223-234, jul./dez., 2001.

ROBREDO, J. **Documentação de hoje e de amanhã.** 4. ed. rev. ampl. Brasília, DF: Ed. Do Autor, 2005.

SAMPIERI, R. H.; COLLADO, C. F..; LUCIO, M. P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SANTOS, H. M. D.; FLORES, D. O documento digital no contexto das funções arquivísticas. **Páginas A&B, Arquivos e Bibliotecas (Portugal)**, n. 5, p. 165-177, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/65458.

SHERA, J. H.; EGAN, M. E. **Examen del estado actual de la biblioteconomía y de la documentación.** Santa Fe, Argentina: Centro de Documentación e Información de Asuntos Municipales Doctor Alcides Greca, 1953.

SILVA, I. C. O.; GOUVEIA, F. C. A busca e o acesso às informações sobre saúde no contexto tecnológico. **Revista Conhecimento em Ação**, Rio de Janeiro, v. 4, n. 2, p. 23-45, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/127448. Acesso em: 25 fev. 2024.

SILVA, M. **Sala de aula interativa.** 4. ed. Rio de Janeiro: Quartet, 2006

SIQUEIRA, J. C. A noção de documento digital: uma abordagem terminológica. **Em Questão**, Porto Alegre, v. 18, n. 1, p. 125–140, 2012. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/24172. Acesso em: 25 fev. 2024.

WEINSTEIN, C; MAYER, R. ***The Teaching of Learning Strategies.*** In: WITTROCK, M. (Ed.) Handbook of Research on Teaching, Macmillan, New York. p. 315-327, 1986.

ZAYAS, E. L. *et al.* **O paradigma da educação continuada.** Porto Alegre: Penso, 2012.