relato de pesquisa

INDEXAÇÃO AUTOMATIZADA DE ARTIGOS DE PERIÓDICOS CIENTÍFICOS: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia

Cristina Miyuki Narukawa *
Isidoro Gil Leiva **
Mariângela Spotti Lopes Fujita ***

Resumo A automatização da indexação tem sido tema de discussões entre pesquisadores da área de Ciência da informação, entretanto são pouco esclarecedoras sobre o uso de software de indexação. Desse modo, verifica-se a necessidade de conhecer os software de indexação, bem como sua aplicação na análise dos conteúdos documentários. Nesse sentido, propôs-se a investigação da consistência na indexação e, da exaustividade e precisão na recuperação da informação mediante análise comparativa entre a indexação automática do Sistema de Indización Semi-Automático (SISA) e a indexação manual do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME). Objetiva-se contribuir para o desenvolvimento teórico da automatização da indexação e aperfeiçoamento do SISA. Para tanto, realizou-se a aplicação e a avaliação do SISA mediante os cálculos dos índices de consistência entre os dois tipos de indexação e os cálculos dos índices de exaustividade e precisão na recuperação por meio de buscas em bases de dados (BDSISA e BDBIREME) constituídas por descritores obtidos pelo SISA e pela indexação manual respectivamente. As variações dos termos utilizados nos artigos científicos em comparação aos que se encontram no DeCS foram os principais fatores dificultadores no alcance de maiores índices de consistência na indexação e refletiu nos índices de exaustividade e precisão na recuperação levando a necessidade de aperfeiçoar a linguagem documentária utilizada no software SISA e a incorporação de métodos lingüísticos.

Palavras-chaves: Indexação automatizada; Programas de indexação, Sistema de Indexação Semi-Automático (SISA), Avaliação da indexação.

* Bolsista da FAPESP no Mestrado do Programa de Pós-Graduação em Ciência da Informação da UNESP – Marília.

** Pesquisador Visitante do CNPq no Programa de Pós-Graduação em Ciência da Informação da UNESP – Marília. Doutor em Filologia Hispanica pela Universidad de Murcia. Professor titular na Universidad de Murcia

*** Doutora em Ciências da Comunicação pela Universidade de São Paulo e Livre Docente em Análise Documentária e Linguagens Documentária Alfabéticas pela Faculdade de Filosofia e Ciências da UNESP – Campus de Marília. Atualmente é Professora Adjunta da Universidade Estadual Paulista Júlio de Mesquita Filho

1 Introdução

Existem muitas discussões na literatura de Ciência da informação referente à automatização do processo de tratamento dos conteúdos temáticos dos documentos, entretanto poucas pesquisas têm esclarecido como se constitui o processo de análise de conteúdo por software de indexação, bem como os aspectos da indexação envolvidos em sua automatização.

A partir do crescimento da produção científica e do desenvolvimento de novas tecnologias em meados do século XX, surgem discussões relacionadas à automatização dos processos de tratamento dos documentos, não apenas para facilitar as tarefas, mas para garantir resultados mais eficazes. Nesse sentido, há esforços para o desenvolvimento de software de indexação, entretanto, é claro que "[...] a indexação automática difere substancialmente

de indexação manual encontrando dessa forma, críticos e defensores". (GUIMARÃES, 2000, p. 1)

Um esforço nesse sentido é o desenvolvimento do Sistema de Indización Semi-Automático (SISA) proposto pelo Prof. Dr. Isidoro Gil Leiva da Universidade de Murcia na Espanha como resultado do seu estudo sobre automatização da indexação. A metodologia aplicada por esse software no processo de análise do documento é efetuada pela comparação entre o documento – constituído por título, resumo e texto – e uma linguagem documentária, a partir de critérios de freqüência preestabelecidos pelo software para propor os termos de indexação.

Diante das discussões sobre a automatização ou não da indexação encontradas na literatura de Ciência da informação e o desconhecimento sobre software de indexação, bem como sua aplicação na análise de conteúdo temático dos documentos, verificamos a necessidade de analisar as características da indexação consideradas importantes na utilização desses software e como é realizada sua aplicação na análise de conteúdo temático dos documentos.

Sendo assim, a pesquisa propôs uma investigação teórica da indexação com ênfase na consistência da indexação e, na exaustividade e precisão na recuperação da informação mediante uma análise comparativa entre a indexação automática de trabalhos científicos, publicados por pesquisadores brasileiros de odontologia, realizada pelo SISA e a indexação manual realizada por bibliotecários que atuam em instituições, centros de documentação ou bibliotecas que participam e cooperam com a Biblioteca Virtual em Saúde (BVS) do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME).

As mudanças nas práticas profissionais influenciadas pelo desenvolvimento das tecnologias de informação e comunicação exigem novos conhecimentos dos bibliotecários. Verifica-se a preocupação em conhecer novos meios de realizar os serviços informacionais diante do grande volume de informações disponíveis atualmente, para permitir que o processo de tratamento do conteúdo temático dos documentos propicie a recuperação de informações relevantes aos usuários. Dessa forma, a pesquisa poderá contribuir para que os sistemas de informação ofereçam aos indivíduos, governos e empresas fácil acesso a essa gama de informações disponíveis nas bases de dados, além de contribuir para que os futuros profissionais conheçam as características e instrumentos auxiliares considerados importantes na elaboração de software de indexação e possam estar preparados para atuar no seu desenvolvimento e uso.

Desse modo, buscou-se realizar o estudo teórico de indexação com ênfase na consistência da indexação e, na exaustividade e precisão na recuperação da informação e analisar o processo de indexação automática mediante análise comparativa da consistência na indexação e da exaustividade e precisão na recuperação da informação entre a indexação automática realizada pelo SISA e a indexação manual de cem documentos científicos de pesquisadores de odontologia realizada por bibliotecários com a finalidade de contribuir para o desenvolvimento teórico da automatização da indexação e aperfeiçoamento do Sistema de Indización Semi-Automático.

2 Fundamentação teórica

As condições recentes do desenvolvimento tecnológico e científico têm gerado uma intensa produção científica e difusão de informações que são elementos fundamentais para o desenvolvimento de novos conhecimentos. Desse modo, verifica-se a necessidade de tratamento adequado do conteúdo temático dos documentos para que as informações sejam recuperadas por aqueles que a necessitam. A forma de tratar o conteúdo dos documentos conhecida como indexação é definida por Pinto (2001, p.226), considerando o raciocínio de Jean-Claude a entrada de ser por Gardin (1974) como,

> [...] um conjunto de atividades que consiste em identificar, nos documentos, os seus traços descritivos (TD's) ou macroproposições e, em seguida, extrair os elementos/descritores (sintagmas) indicadores do seu conteúdo, visando à sua recuperação posterior. Esses descritores vão se constituir na representação dos elementos indicadores do conteúdo do documento e não a sua representação, pois esta só pode ser pelo próprio documento.

A qualidade que se reflete no atendimento das necessidades informacionais dos usuários está diretamente relacionada às características

inerentes ao processo de indexação, entre as quais se encontram: a exaustividade, a especificidade, a correção e a consistência na indexação. Essas características da indexação são determinadas a partir de decisões políticas da organização, portanto, espera-se que ao projetar e utilizar um software de indexação essas características também sejam contempladas, visto que são essenciais no contexto da recuperação da informação.

2.1 Automatização da indexação

Com o objetivo de verificar o que efetivamente tem sido concretizado em relação à automatização da indexação e compreender como é efetuada a análise de conteúdo pelos software de indexação, além de verificar a posição e argumentos de pesquisadores, será abordada a automatização da indexação apresentando os conceitos, argumentos contra e a favor, sua evolução e métodos e, apresentando também alguns software encontrados na literatura de Ciência da Informação com enfoque no SISA.

A análise automática de textos é considerada uma área de pesquisa importante a mais de trinta anos e, continua sendo de grande interesse na Ciência da informação, visto que atualmente existe a preocupação de oferecer acesso mais rapidamente à literatura técnico-científica e é possível utilizar o computador no processamento de dados e informações (ROBREDO, 2005). De acordo com Robredo (2005, p.170) a indexação automática consiste em "[...] qualquer procedimento que permita identificar e selecionar os termos que representam o conteúdo dos documentos, sem a intervenção direta do indexador". Acrescenta que no processo de indexação automática, um algoritmo, em certa medida, realiza a tarefa do indexador na escolha dos termos significativos.

Na revisão de literatura sobre a automatização da indexação realizada por Gil Leiva (1999, p. 57; 2008, p. 320) foram identificadas vinte expressões diferentes sobre automatização da indexação que, em realidade, se referem a três conceitos diferentes:

a) aquele em que os programas realizam o processo de armazenamento dos termos de indexação obtido por um profissional, chamado de indexação assistida por computador;
b) os programas realizam a análise dos documentos de modo automático e se necessário os termos são validados por um profissional, denominado indexação semi-automática; e
c) os programas realizam o processo de análise dos documentos e não ocorre validação por profissionais, ou, a indexação automática propriamente dita.

Embora se proponha a aplicação e análise do sistema semi-automático de indexação – o SISA –, é importante destacar que será abordado o conceito de indexação automática, já que é imprescindível compreender os aspectos envolvidos no processo automático do SISA, no qual se pretende dar enfoque ao realizar a comparação com a indexação manual.

Sob o ponto de vista do modo com que os termos de indexação são selecionados Lancaster (2004) distingue dois tipos de indexação, a indexação por extração automática em que as palavras ou expressões do texto são extraídas e utilizadas para representar o texto como um todo, ou seja, a indexação é realizada a partir da linguagem natural. E a indexação por atribuição automática que consiste na representação do conteúdo mediante termos selecionados de alguma linguagem documentária, o que segundo Lancaster (2004, p. 289) é considerada mais difícil quando aplicada a computadores e é necessário "[...] desenvolver, para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer freqüentemente nos documentos [...]".

Foram identificadas na literatura de Ciência da informação, discussões relacionadas às vantagens e desvantagens da aplicação de software de indexação. Gil Leiva (2008, p. 320) sistematizou esses argumentos, os contra e a favor da automatização da indexação e verificou que os argumentos a favor da automatização da indexação:

- Caracteriza a indexação humana como lenta, subjetiva e de alto custo;
- Afirmam que com a automatização da indexação tem-se diminuição de erros que repercutirá positivamente na recuperação das informações em bases de dados;

- A indexação parece ser mais precisa, permitindo uma recuperação dos documentos mais rica.
- Enquanto os argumentos contra:
- Alegam sobre a incapacidade dos sistemas automáticos de indexação reconhecerem diferentes significados em diferentes contextos, relacionar e selecionar conceitos implícitos dos documentos;
- Afirmam que a indexação automática reconhece palavras e não conceitos, defendem que se deve perseguir a captação de terminologias dos textos, porque esta cumpre a função representativa, cognitiva e comunicativa que apresentam os conceitos e, por tanto, o conhecimento;
- Alegam que na maioria das ocasiões a automatização da indexação restringe-se às áreas específicas do conhecimento;
- Uma das principais razões abordadas contra a automatização do processo é a impossibilidade, no estado atual da investigação, conseguir indexação totalmente automática.

Nesse sentido, de acordo com Silva e Fujita (2004), o problema da indexação automática é que os assuntos dos documentos não são representados da mesma forma que a indexação humana o faz, e acrescenta que isso se deve ao fato de que ainda se desconhece o processo mental envolvido na análise de assunto durante o processo de indexação, deste modo, enquanto não se conhecer tais processos não será possível atribuir indexação semelhante aos computadores.

De acordo com Moreiro González (2004) os procedimentos da indexação automática e indexação manual não são equivalentes, isso porque não é possível que as máquinas imitem a capacidade humana e, diz ainda que estaríamos cometendo um erro grave se pretendêssemos isso. Não se trata de justificar a automatização ou não da indexação, mas nas atuais circunstâncias de crescimento informativo, a questão se centra na necessidade de criar um *software* eficaz que automatize o processo, levando-se em conta que os documentos indexados de maneira automática respondem a padrões determinados e que a indexação automática não poderá dar conta de alguns aspectos que só podem ser obtidos mediante a análise humana.

Lancaster (2004) explica que embora se fale em índices razoáveis de desempenho da indexação automática, geralmente não alcançam o nível de desempenho dos indexadores humanos, entretanto, esse tipo de indexação poderá reduzir a carga de trabalho dos indexadores humanos ao realizar a indexação preliminar.

A posição de Lancaster (2004) e Moreiro González (2004) condiz com a proposta dos software de indexação semi-automáticos na medida em que defendem o uso de *software* para indexar documentos com a ressalva de que é necessário ter em mente as limitações em seu processo, as quais só a análise apurada de um ser humano pode solucionar, sendo clara a necessidade de avaliação final dos termos de indexação por um profissional.

A aplicação da automatização da indexação tem se desenvolvido como alternativa ao tratamento da informação diante do crescimento exponencial de documentos, situação que Robredo (2005) expõe ao dizer que a necessidade de indexar grandes volumes de informações, em um tempo curto para manter as bases de dados atualizadas, tornou inviável pensar na indexação manual (humana ou intelectual) como única forma de analisar e codificar o conteúdo dos documentos. Dessa forma, Robredo (2005) compreende que as pesquisas relacionadas à indexação automática devem se desenvolver ao mesmo tempo em que as pesquisas em indexação manual (humana ou intelectual).

As pesquisas sobre a automatização da indexação iniciam-se na segunda metade do século XX. Para compreender melhor como ocorreu o desenvolvimento da automatização da indexação, e verificando que a literatura de Ciência da informação apresenta esse processo em três momentos históricos diretamente relacionados, partimos desse princípio para esquematizar sua evolução.

2.2 Evolução da automatização da indexação

De acordo com Gil Leiva (1999, 2008) o desenvolvimento da automatização da indexação ocorreu em três momentos históricos: o dos métodos estatísticos, o dos métodos lingüísticos e o dos métodos mistos ou híbridos.

2.2.1 Métodos estatísticos

Para construção de índices, Hans Peter Luhn confrontava o documento com uma lista de palavras vazias, no entanto, era necessário aplicar um critério que determinasse as palavras significativas dentro daquele conjunto de palavras restantes, concedendo maior peso a umas do que outras. Esses novos critérios foram desenvolvidos calculando-se a freqüência estatística de aparição das palavras. Segundo Vieira (1988) o método de freqüência consiste na contagem automática do aparecimento da palavra, que pode estar no título, resumo, título das referências citadas, texto e em diversas combinações dessas fontes.

A utilização do critério de freqüência só foi possível através dos métodos estatísticos que foram os primeiros a surgir. Zipf em 1949 desenvolveu o princípio do mínimo esforço que se referia ao valor constante que tem a relação entre a freqüência das palavras e a posição que essas ocupam na ordem frequencial. A partir dessas idéias, Hans Peter Luhn em 1957 sugeriu que a freqüência das palavras em um texto tem relação com a utilidade que teriam na indexação. Para ele, a freqüência com que as palavras aparecem no texto expressa quais são as palavras representativas do conteúdo do texto. Essas idéias permitiram o surgimento da indexação ponderada que consiste em atribuir um valor de importância aos termos de indexação com relação ao documento indexado. Segundo Gil Leiva (1999, 2008) outros métodos de ponderação dos termos surgiram, tais como: o valor de discriminação dos termos proposta por Salton e Yang (1973), a função de freqüência inversa proposta por Sparck Jones (1972) e aplicações de método de freqüência na coleção, como em Damerau (1965), que comparou a freqüência de uma palavra em um documento com a freqüência da mesma palavra na coleção para determinar sua inclusão ou não como termo de indexação.

Entretanto, os métodos puramente estatísticos apresentam alguns problemas para a automatização da indexação relacionados aos aspectos lingüísticos do texto.

2.2.2 Métodos lingüísticos

Diante das limitações dos métodos estatísticos em relação aos aspectos lingüísticos do texto, buscou-se introduzir critérios de discriminação lingüística das palavras.

Nesse sentido, Gil Leiva (1999, p. 82) explica que a partir do início dos anos sessenta associam-se as técnicas de Processamento da Linguagem Natural (PLN) que consiste no estudo e análise dos aspectos lingüísticos de um texto mediante a utilização de programas informáticos – e a automatização da indexação. Os estudos lingüísticos avançaram em direção à compreensão da estrutura textual, suas relações e seu significado.

Segundo Gil Leiva (2008, p. 339) os primeiros analisadores lingüísticos surgiram na década de 1960 para o processamento automático de informação. Os avanços e melhorias produzidas nestes sistemas têm permitido utilizá-los para a recuperação da informação, extração de informação, classificação, indexação e resumos de documentos ou para o reconhecimento automático da fala. Estes analisadores lingüísticos (Figura 1), coincidindo com os níveis de linguagem, se dedicam ao tratamento das palavras (analisador morfológico), ao tratamento das orações (analisador sintático) e ao tratamento das palavras e orações segundo o contexto em que se encontram para conhecer seu significado (analisadores semânticos).

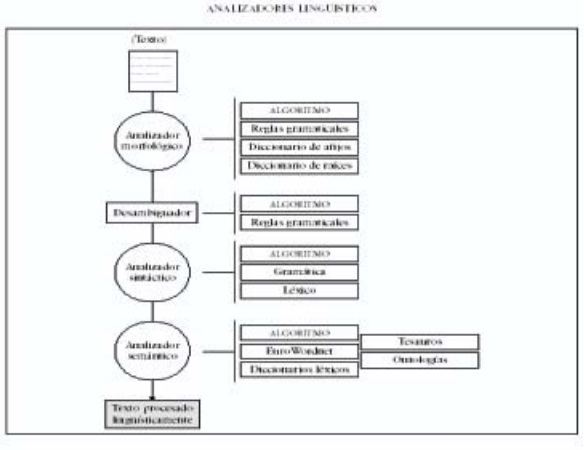

Figura 1: Analisadores lingüísticos
Fonte: GIL LEIVA, 2008, p. 339

A seguir, resumimos em um quadro alguns aspectos relativos a estes três tipos de analisadores lingüísticos:

MORFOLÓGICO	A *Morfologia* é a parte da Lingüística que estuda as formas das palavras e identifica seus elementos constitutivos. Elementos importantes das palavras são as raízes (parte que aporta significado) e os afixos (acompanham a raiz como prefixos, que precedem a raiz como ex; -ante; -re, etc., ou sufixos, que seguem a raiz como –o; -a; -as; -e; -ível; -ando; -mente; -ista; -ismo; etc.). A Morfologia também se dedica ao estudo da combinação destes elementos para produzir novas palavras mediante processos de flexão (janelas > janela + s; o s indica plural) e derivação (armazém + agem = armazenagem). A lematização (*stemming* em inglês) é a redução de uma palavra ou conjunto de palavras a sua raiz uma vez detectadas ou eliminadas suas variantes flexivas (número, gênero, desinência) e derivativas (-istico; -avel; -dade; -ista; -ção; etc.) mediante um programa informático. A utilidade da lematização é dupla: 1º. No tratamento da informação. Para que no cálculo da freqüência de aparição de um termo executado durante a indexação automática, os termos que possuem uma mesma raiz se computem conjuntamente.
SINTÁTICO	A sintaxe é a parte da gramática que ensina a coordenar as palavras para formar as orações e expressar conceitos. Considerando como início o resultado anterior obtido pelo analisador morfológico, o analisador sintático desambigua o resultado do morfológico para estabelecer a categoria gramatical de cada uma das palavras e as relações entre elas, por meio de alguma representação estruturada, normalmente arbórea. Os analisadores sintáticos (parking) se apóiam em três ferramentas: uma gramática, um dicionário com as palavras e suas possíveis categorias gramaticais ou o resultado de um analisador morfológico e um algoritmo de análise.
SEMANTICO	A semântica estuda a significação das palavras em relação a outras e por extensão, o significado das palavras em um texto ou em um diálogo. Como é sabido, uma característica das línguas é sua riqueza semântica por emprego consciente ou não de recursos como a sinonímia, a polissemia, a anáfora ou a elipse. Os analisadores semânticos são ferramentas desenvolvidas para resolver de maneira automática alguns tipos de sinonímia, polissemia, anáfora ou elipse, já que a resolução ou desambiguação integral e conjunta destes recursos lingüísticos é na atualidade tremendamente difícil por sua complexidade e o grande número de conhecimento lingüístico e de programação requeridos.

Quadro 1: Aspectos dos analisadores lingüísticos

Fonte: Gil Leiva (2008, p.339-349).

Com a aplicação dos métodos lingüísticos ao processamento de textos houve grande avanço da automatização da indexação. Para o desenvolvimento da indexação automática é extremamente importante que as pesquisas e aplicações direcionem esforços para solucionar os problemas de processamento lingüístico, visto que grande parte dos problemas enfrentados na indexação automática é de base terminológica e lingüística e, é claro aliar o conhecimento de outras áreas relacionadas.

2.2.3 Métodos mistos ou híbridos

Na atualidade, segundo Guimarães (2000) verificam-se os métodos mistos ou híbridos

de indexação automática que reúnem aportes da estatística, da lingüística textual e ainda utilizam tesauros como instrumento de controle de vocabulário, auxiliando e contribuindo para eliminar problemas como a sinonímia e a identificação de funções sintáticas dos termos, proporcionando benefícios à revocação na recuperação da informação.

As últimas tendências da automatização da indexação é denominada de indexação inteligente por Mendez Rodríguez e Moreiro González (1999). Explicam que esse tipo de indexação está voltado ao acesso direto de documentos por meio do processamento lingüístico automático e uso de linguagem natural combinando outras técnicas como análise estatística ou a ponderação dos termos. Esses sistemas buscam interfaces inteligentes para que o usuário possa utilizar a linguagem natural como linguagem de intercâmbio de conhecimento e é atribuída ao computador a competência lingüística e/ou cognitiva, tendo não só bases lingüísticas, mas também bases de conhecimento.

2.3 Software de indexação

Os software de indexação apresentados a seguir são encontrados na literatura de Ciência da informação. No Brasil são identificadas poucas iniciativas nesse sentido, com destaque para o software Automindex descrito por Robredo (1991) e no âmbito internacional é possível encontrar uma literatura consideravelmente vasta sobre o assunto, já que como aponta Gil Leiva (2003) os avanços na indexação automática têm sido aplicado em determinadas unidades documentais que operam grande quantidade de informação, e, portanto, é necessário automatizar, na medida do possível, os processos de análise e tratamento para agilizar os processos. Deste modo, surgiram protótipos como "Shapire" desenvolvido pela Biblioteca Nacional de Medicina dos Estados Unidos (HERSH; GREENES, 1990); no centro de documentação da NASA (SILVESTRE; GENUARDI; KLINGBIEL, 1994); ou mais recentemente, no Laboratório Europeu de Física de Partículas (CERN) de Genebra (MONTEJO RÁEZ, 2001), entre outros.

Ainda que não exista uma metodologia única para sua projeção e desenvolvimento; ou de protótipos de indexação automática, quase todos os sistemas contêm algum dos elementos sistematizados por Gil Leiva na figura 2 (2008, p. 367):

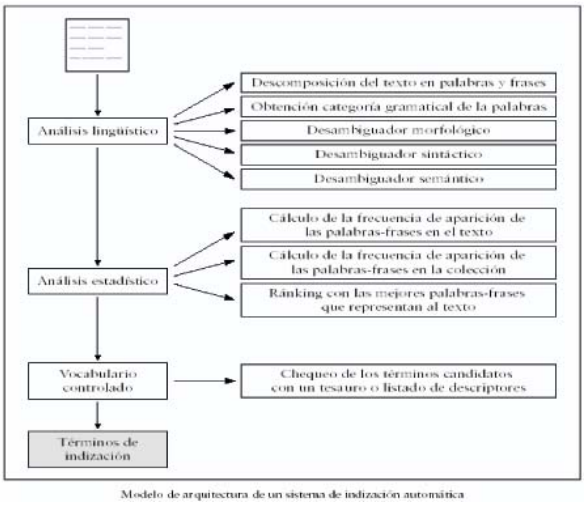

Figura 2: Modelo de arquitetura de um sistema de indexação automática

Fonte: GIL LEIVA, 2008, p.367

Aqui destacamos o software Automindex, como iniciativa nacional, e o SISA por sua importância para pesquisa.

O sistema de indexação Automindex apresentado por Robredo (1991) possui como uma de suas características principais a existência de dois antidicionários concomitantes de palavras vazias: um de invariáveis, tais como conectivos, e outro de raízes de palavras não significativas para a área de conhecimento considerada. As vantagens de se utilizar esses dois dicionários são: a diminuição do volume total do dicionário, a diminuição do volume de memória necessário para armazenagem do dicionário, a diminuição do volume necessário para processamento e o aumento da velocidade de processamento. Robredo (1991) explica que para o processamento são considerados os títulos e os resumos, sendo que os caracteres a serem analisados são delimitados previamente.

Em um primeiro momento o texto é analisado comparando-se as palavras do texto com as do dicionário de invariáveis, se forem identificadas no dicionário, são desprezadas. O mesmo ocorre no processo seguinte com as palavras comparadas com o dicionário de raízes de palavras significativas. As palavras que restaram são consideradas como possíveis descritores. Para considerar como descritores as palavras são comparadas com um dicionário de palavras significativas, se identificadas são descritores e as

que não forem identificadas como significativas são consideradas "candidatas a descritores" para exame e avaliação que decidirá se será incorporada ou não. Por fim, é possível listar os descritores e candidatos a descritores com suas respectivas freqüências de aparecimento na base de dados.

O Automindex revela-se versátil para indexação tanto de títulos e resumos de documentos para a geração de índices temáticos e organização de base de dados para recuperação em linha, como para a indexação de documentos correntes em arquivos empresariais, tais como cartas, ofícios, etc. (ROBREDO, 1991).

O SISA (Sistema de Indización Semi-Automático) desenvolvido na Espanha por Gil Leiva (1999, 2008), foi inicialmente proposto para a área de Biblioteconomia e Documentação, no entanto, a flexibilidade do sistema permite adaptar sua configuração para aplicar a qualquer área, desde que possua uma linguagem documentária. O SISA é um sistema semi-automático de indexação, conforme Diagrama de fluxos do algoritmo SISA apresentado na figura 3, que analisa as partes do documento que estão delimitados com marcadores para que o sistema possa reconhecer as fontes (título, resumo e texto) e aplicar seus critérios para propor os termos de indexação.

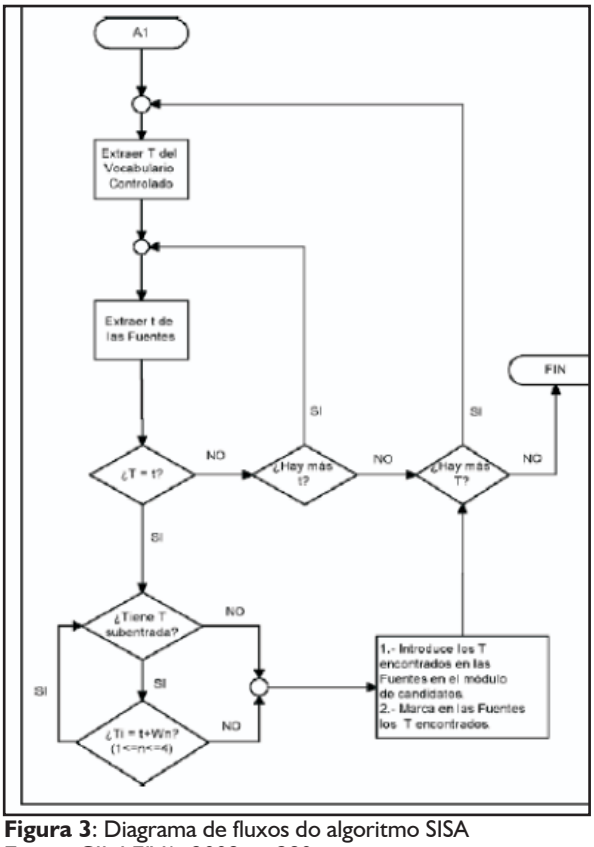

Figura 3: Diagrama de fluxos do algoritmo SISA
Fonte: GIL LEIVA, 2008, p. 380

As fontes utilizadas no processamento pelo SISA são: o texto completo (título, resumo e texto), uma lista de palavras vazias (*stoplist*) e uma linguagem documentária, todos em formato txt.

O processo de indexação se desenvolve em três módulos:

O módulo 1 é o pré-processamento em que o documento é preparado sinalizando-se as partes com os marcadores #CTI# (começo do título), #FTI (fim do título), #CR# (começo do resumo), #FR# (fim do resumo), #CTE# (começo do texto) e #FTE# (fim do texto). Além disso, as frases e orações compreendidas entre sinais de pontuação são horizontalizadas, ocorre também a eliminação das palavras vazias mediante a comparação com a lista de palavras vazias e então é computado o total de palavras das fontes título, resumo e texto.

Em um segundo momento, *no módulo 2* ocorre a etapa de análise do conteúdo, processamento em que um algoritmo busca e seleciona os termos preferidos que são os coincidentes com os termos da linguagem documentária; os termos não preferidos que são os termos sinônimos, por isso não podem ser utilizados e remetem aos preferidos e; os termos construídos sintaticamente de forma diferente dos termos preferidos que são as palavras semi vazias, aquelas que o sistema julga importante, mas não se enquadram nas anteriores.

O terceiro e último módulo é a etapa de valoração e ponderação que consiste na aplicação de um critério de avaliação dos termos para que o sistema possa selecionar os termos de indexação que representarão o conteúdo do documento. Isso é necessário, pois do contrário, o sistema selecionaria todos os termos da linguagem documentária que coincidem com os das fontes.

Para selecionar e propor os termos para indexação são aplicados os seguintes critérios: o termo é apresentado como termo de indexação se, um termo autorizado aparece na fonte título e na fonte resumo, ou se um termo autorizado aparece na fonte título e na fonte texto, ou se um termo autorizado aparece na fonte resumo e na fonte texto. No entanto, os termos não autorizados ou semi vazios são apresentados como candidatos à indexação se, a palavra semi vazia aparece no título, resumo e texto, ou se aparece no texto dez vezes ou mais,

além de aparecer em oito parágrafos diferentes ou mais.

A etapa posterior para concluir a indexação já não depende tanto do sistema, mas conta com a decisão do indexador humano que deverá analisar os termos de indexação propostos pelo sistema e os termos semi vazios candidatos a descritores que se apresentam em uma lista e podem ou não ser incluídos como termos de indexação, além do sistema oferecer a possibilidade de acrescentar termos da linguagem documentária. O sistema permite que a etapa seja flexível, sendo possível acrescentar ou suprimir termos, isso como forma de permitir que o indexador possa tomar a decisão considerando as particularidades do sistema de informação.

Em seguida, se apresenta a interface do SISA (FIGURA 4):

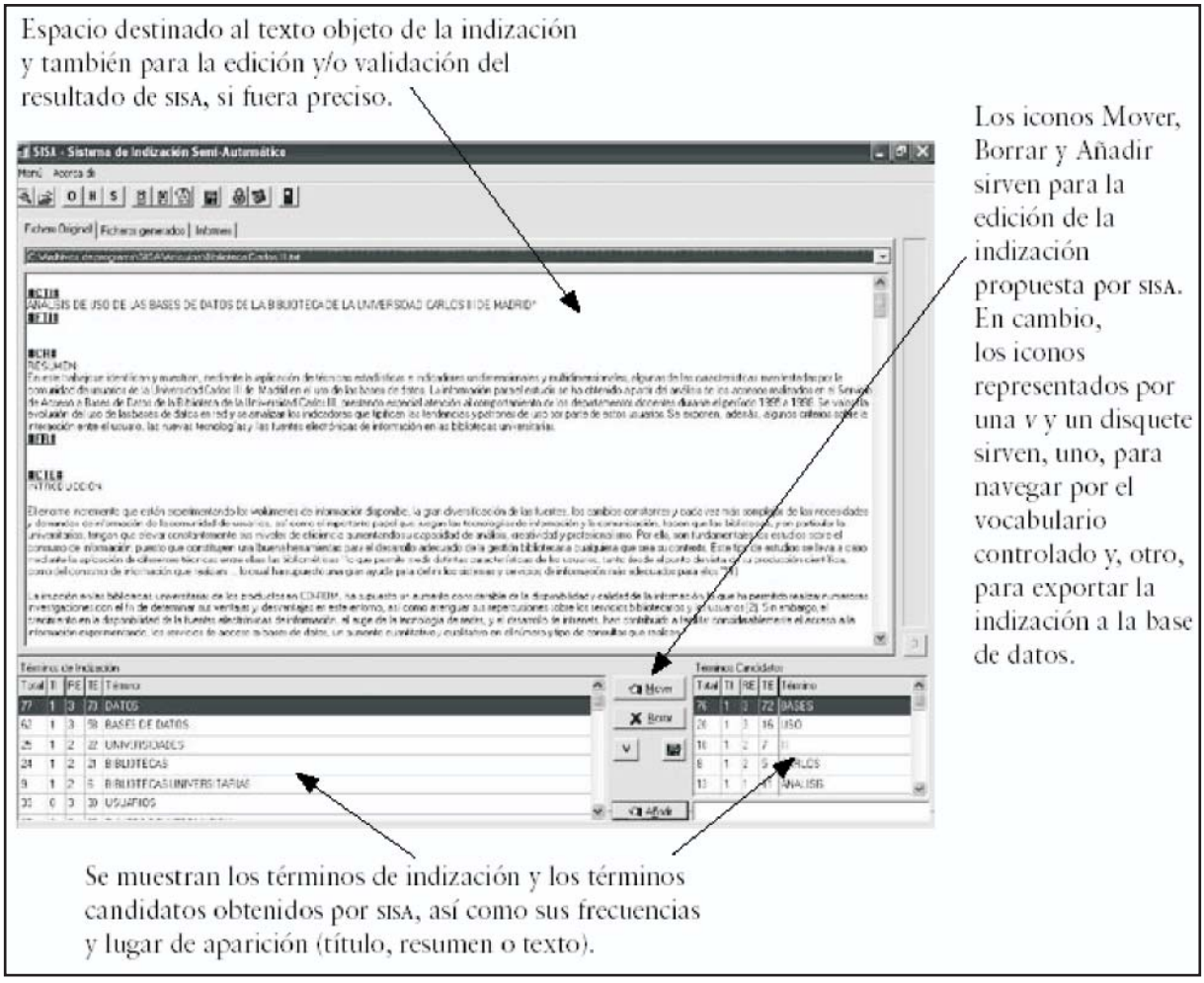

Figura 4: Interface do SISA
Fonte: Gil Leiva, 2008, p. 383

3 Avaliação da indexação

Segundo Gil Leiva (2008, p. 38) a avaliação da indexação pode ser intrínseca ou extrínseca. Vejamos de forma resumida cada uma destas formas de avaliação propostas pelo autor:

3.1 Avaliação intrínseca da indexação

Intrínseco significa interno, íntimo, portanto, a avaliação intrínseca da indexação é o conjunto de tarefas centradas no resultado da indexação (descritores, cabeçalhos, subcabeçalhos

ou identificadores) com a finalidade de conhecer sua qualidade. A avaliação intrínseca da indexação pode ser qualitativa, isto é, por meio de valoração e consensos entre especialistas ou quantitativa através de fórmulas.

3.1.1 Avaliação intrínseca qualitativa

Na avaliação intrínseca qualitativa se analisa cada um dos componentes que conjuntamente proporciona a qualidade da indexação. A análise destes elementos se conduz por pelo menos dois especialistas que devem conhecer a política de indexação, a linguagem de indexação, as características dos usuários do sistema de informação, etc.

Segundo Gil Leiva (2008) o procedimento para executá-la é o seguinte: a) os especialistas selecionam aleatoriamente um número significativo de registros da base de dados; b) em seguida, ou a indexação de cada documento é realizada novamente considerando como referência o texto completo do mesmo, o que implica uma dedicação e tempos consideráveis, ou somente se estuda com detalhamento a informação do registro selecionado, o que permite uma revisão da indexação baseada no título e também no resumo do documento. Independente do procedimento seguido, os especialistas devem emitir valorações e conseguir consensos ao menos nestes componentes inerentes a uma indexação de qualidade: Exaustividade (Que se extrai todos os conceitos caracterizadores do conteúdo íntegro dos documentos), Especificidade (Que exista uma relação exata entre as unidades conceituais escolhidas e o termo ou os termos eleitos para representá-la), Correção (Que não se produza erros de inclusão (um termo que não corresponde) nem erros de omissão (a exclusão de um termo que corresponde) e Perspectiva do usuário (Que se considere os interesses e necessidades do usuário).

3.1.2 Avaliação intrínseca quantitativa

Por outro lado, a avaliação intrínseca quantitativa é a reindexação de um conjunto de documentos repetindo, na medida do possível, o ambiente em que se produziu a primeira indexação (indexadores, política de indexação, linguagem de indexação, condições de trabalho, interesses e necessidades dos usuários potenciais, etc.) para estabelecer comparações entre dois indexadores. Portanto, o grau de consistência será maior quanto mais similares forem as indexações comparadas. Esta semelhança ou diferença entre indexadores pode ser quantificada através de fórmulas.

As equações para encontrar os índices de consistência entre dois indexadores (QUADRO 2) são as seguintes:

Hooper (1965)	Rolling (1981)
$\frac{C}{A+B-C}$ Uma variação desta equação é: $\frac{100C}{C+A+B}$ onde, C= Termos comuns nas duas indexações A= Termos usados na indexação A mas não em B B= Termos usados na indexação B mas não em A	$\frac{2C}{A+B}$ onde, C= Termos comuns nas duas indexações A= Termos usados na indexação A B= Termos usados na indexação B

Quadro 2: Equações de índices de consistência

Fonte: Gil Leiva (2008, p.386)

A fórmula de Hooper vem sendo expressa como índice de consistência por Gil Leiva (1999, 2003 e 2008) da seguinte maneira:

$$C_i = \frac{T_{co}}{(A+B) - T_{co}}$$

onde,

T_{co} = Número de termos comuns nas duas indexações

A= Número de termos usados na indexação A

B= Número de termos usados na indexação B

Como se pode verificar, com esta equação os índices de consistência oscilam entre os valores 0 e 1 e depois é possível multiplicar o resultado por cem para obter o percentual. Vejamos em dois exemplos, a atribuição de assuntos de um livro e a indexação de um artigo:

Assuntos atribuídos a um livro

Indización A	**Indización B**
1. Drogas-Tráfico	1. Cocaína-Consumo 2. Cocaína-Tráfico 3.Cocaína-Aspectos políticos 4. Cocaína-Aspectos sociais
$C_i = \frac{1}{1+4-1}$ = 0,25 x 100 = 25 %	

Descritores atribuídos a um artigo

Indización A	**Indización B**
1. Mercado de trabalho 2. Ofertas de emprego 3. Diários 4. Documentação 5. Biblioteconomía 6. Arquivística 7. Documentalistas 8. Bibliotecários 9. Arquivista	1. Documentalistas 2. Bibliotecários 3. Arquivista 4. Ofertas de emprego 5. Mercado de trabalho 6. Diários 7. Requisitos professionais
$C_i = \frac{6}{9+7-6}$ = 0,6 = 60%	

Quadro 3: Aplicação do índice de consistência.
Fonte: GIL LEIVA, 2008, p.387

A avaliação intrínseca quantitativa é de grande utilidade para avaliações periódicas em uma mesma unidade de informação por meio de experimentos de intraconsistência, isto é, quando um profissional indexa novamente um documento transcorrido um tempo (seis ou doze meses) para comprovar se se produzem variações com respeito à primeira indexação.

3.2 Avaliação extrínseca

Gil Leiva (2008 p.388) apontou que na avaliação intrínseca da indexação se realiza uma análise interior enquanto que na avaliação extrínseca o resultado da indexação é utilizado para compará-lo com a indexação de outra unidade de informação que também indexou os mesmos documentos (interconsistência) ou para testar a função da indexação na recuperação (exaustividade e precisão na recuperação).

3.2.1 Avaliação extrínseca mediante a interconsistência

Neste caso, se aplica igualmente a fórmula de consistência estudada acima, mas aqui se compara as indexações de duas instituições (Biblioteca A e Biblioteca B) ou sistemas de indexação (uma indexação manual e outra automática) que tem indexado o mesmo documento, e se possível com a mesma ferramenta de indexação.

Na comparação da indexação de duas instituições, os elementos a serem

considerados na comparação de indexações devem estar presididos por uma mínima homogeneidade entre os fatores que afetam a indexação. Gil Leiva continua, apontando que os fatores que mais intervêm no resultado da indexação são o próprio indexador, o objeto analisado e o contexto em que se concretiza. Assim, os *fatores relativos ao indexador* têm a ver com a formação e experiência em indexação, o conhecimento do assunto, o domínio da ferramenta de indexação ou a profissionalismo; os *fatores relativos ao contexto* com a política de indexação definida pela instituição, os tipos e necessidades dos usuários ou a carga de trabalho e o tempo dedicado; e os *fatores relativos ao objeto* analisado com sua complexidade, suas características e propriedades, tamanho, etc.

3.2.2 Avaliação extrínseca mediante a recuperação

Esta avaliação serve para comparar dois indexadores procedentes do mesmo sistema (Intraconsistência) ou dois sistemas de informação diferentes, neste caso pode ser uma indexação manual com outra automática, duas automáticas ou as duas manuais.

A avaliação extrínseca mediante a recuperação consiste em interrogar duas bases de dados que contém os mesmos campos e idênticos conteúdos salvo os campos que armazenam a indexação. Com os resultados obtidos se encontram os índices de exaustividade e precisão na recuperação. Este método de avaliação é custoso, mas proporciona resultados claros, mediante o uso de fórmulas de exaustividade e precisão na recuperação:

$$Exaustividade = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ de\ documentos\ relevantes\ na\ coleção}$$

$$Precisão = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ recuperados}$$

No Quadro 4 resumimos o procedimento para uma avaliação extrínseca mediante a recuperação de informação:

1. Construir duas bases de dados com no mínimo cem registros. As duas bases de dados devem possuir os mesmos campos e idêntica informação em cada um deles exceto aqueles que contêm o produto da indexação. Assim, uma base de dados terá uma indexação A, que será a indexação do sistema de informação ou indexador A, enquanto a indexação B corresponderá ao sistema de informação ou indexador B.
2. Atribuir a cada um dos documentos incluídos nas duas bases de dados sua relevância temática. Isto significa estabelecer para quais questões um determinado documento é relevante. Exemplos:

Documentos
Relevantes para

Documento 1 :
Osteoma ; Diagnóstico por imagen

Documento 2 :
Hiperplasia gengival ; Preparaçoes farmacêuticas ; Placa dentária

Documento 3 :
Cistos odontogênicos ; Cisto periodontal
…
…

3. Selecionar um conjunto de necessidades de informação reais que tenham relação com o conteúdo das bases de dados. Desta maneira, sabemos quais documentos da base de dados são relevantes para cada uma das petições de informação selecionadas. Este controle nos permite usar fórmulas de exaustividade e precisão na recuperação mostrada anteriormente. E é justamente, esta necessidade de controlar os documentos relevantes o que impossibilita que este tipo de avaliação se execute em base de dados com milhares de documentos, por isso se recorre à *experimentação de laboratório* com centenas de documentos como máximo.

4. Construir as equações de interrogação para cada petição de acordo com os parâmetros próprios do sistema e iniciar as buscas.

5. Anotar o número de cada um dos documentos recuperados para cada uma das necessidades de informação.

6. Encontrar os índices de exaustividade e precisão para cada uma das petições utilizando para tanto as fórmulas correspondentes, as relevâncias temáticas conseguidas no passo 2, assim como a resposta oferecida pelo sistema.

7. Encontrar a exaustividade e precisão média para cada uma das bases de dados.
Por fim, a maior ou menor exaustividade e precisão obtida são devido a indexação, entendendo que os outros dados dos registros são os mesmos nas duas bases de dados. Assim, a base de dados que consegue melhores índices de exaustividade e precisão na recuperação significa que possui uma melhor indexação.

Quadro 4: Procedimento para avaliação extrínseca por meio da recuperação.
Fonte: Gil Leiva (2008, p. 392)

4 Avaliação do SISA

Após revisar algumas questões básicas para a avaliação da indexação, vejamos como se realizou a avaliação do software de indexação SISA.

Para realizar a avaliação do SISA efetuou-se a análise comparativa entre indexação automática do SISA e indexação manual da BIREME. Para tanto, foi necessária a aplicação do SISA que consistiu em três fases: preparação das fontes a serem utilizadas na indexação, testes e a fase de indexação definitiva dos artigos científicos.

a) Preparação das fontes

LISTA DE DESCRITORES: organizou-se a linguagem documentária DeCS de odontologia estruturando os descritores, com suas respectivas remissivas, em ordem alfabética, enumerando-os e convertendo o arquivo em txt., como exige o SISA. O uso da linguagem documentária DeCS foi autorizado pela BIREME que atendeu à solicitação da lista de termos da área de odontologia.

LISTA DE PALAVRAS VAZIAS: uma lista de palavras vazias em língua portuguesa foi constituída por termos organizados em ordem alfabética e o arquivo convertido em txt.

ARTIGOS CIENTÍFICOS: cem artigos científicos de pesquisadores brasileiros de odontologia publicados na Revista Odonto Ciência, do período de 2004 a 2007 e da Revista de Patologia Oral, do período de 2004 a 2005 que se encontram na base de dados BBO na BIREME foram selecionados e estruturados com marcadores de título, resumo, texto e, convertidos em formato txt. Esses artigos já se encontravam indexados manualmente com o uso da linguagem documentária DeCS por bibliotecários (indexadores experientes) de instituições que participam e cooperam com a BIREME.

b) Testes

Foram realizados dois testes com apenas vinte artigos científicos para verificar a adequação das fontes utilizadas no SISA. O primeiro teste revelou a necessidade de adequar a linguagem documentária com o acréscimo de descritores de odontologia. Após o acréscimo realizou-se novo teste que demonstrou adequação das fontes.

c) Indexação definitiva dos artigos científicos

Após os testes, foi realizada a indexação dos cem artigos científicos pelo SISA com os seguintes procedimentos:

- Configuração no SISA: a partir do SISA as fontes "Lista de palavras vazias" e "Lista de descritores" foram selecionadas, inseridas e armazenadas no SISA;
- Seleção dos artigos científicos no SISA: a partir do SISA, dez artigos científicos foram selecionados, inseridos e armazenados no SISA;
- Prosseguiu-se a execução da indexação automática dos artigos científicos que foram selecionados;
- Os termos de indexação e os termos candidatos a indexação propostos pelo SISA foram exibidos pelo software e;
- Prosseguiu-se novamente a etapa de seleção dos artigos científicos para indexar os demais artigos científicos.

Após obter os descritores dos cem artigos científicos seguiu-se a fase de avaliação desses descritores que será apresentada à seguir.

Os descritores obtidos pelo SISA e os obtidos pela BIREME foram organizados em um quadro comparativo. Através do quadro comparativo foi possível verificar quais e quantos descritores iguais aos descritores da BIREME foram atribuídos pelo SISA, permitindo a execução da próxima etapa de avaliação que propõe verificar a consistência da indexação.

4.1 Avaliação da consistência da indexação com o sisa

Para calcular os índices de consistência em relação aos termos coincidentes determinou-se que quando houvesse coincidência total entre um termo do SISA e um termo da BIREME, ou diante de dois sinônimos atribuir-se-ia 1 e quando houvesse coincidência parcial atribuir-se-ia 0,5. O estudo considerou a indexação realizada por profissionais da BIREME como parâmetro de qualidade para avaliar a indexação do SISA, ou seja, os índices de consistência poderiam oscilar entre 0 a 100% e quanto mais próxima a indexação do SISA estivesse da indexação da

BIREME, consideraríamos maior a qualidade dos descritores propostos por este software de indexação.

4.2 Avaliação da exaustividade e precisão na recuperação com o sisa

Após avaliar o SISA com relação à consistência na indexação, seguiu-se a avaliação da exaustividade e precisão na recuperação da informação por meio da efetuação de buscas em bases de dados constituídas pelos descritores dos dois tipos de indexação. A base de dados BDSISA foi constituída pelos registros dos cem artigos científicos e seus respectivos descritores propostos pela indexação automática do SISA e a base de dados BDBIREME com os registros desses mesmos artigos científicos, mas com os descritores obtidos pela indexação manual da BIREME.

Para verificar se os descritores são eficazes na recuperação das informações, foram estabelecidas previamente as necessidades de informação que seriam apresentadas nas buscas, estabelecendo também os respectivos artigos científicos relevantes, ou seja, que respondem a essas necessidades de informação. Dessa forma, foram realizadas cinqüenta buscas em todas as bases de dados e os resultados foram submetidos aos cálculos dos índices de exaustividade e precisão na recuperação da informação.

O índice de exaustividade na recuperação é obtido através da relação entre os artigos científicos relevantes recuperados e o total de artigos científicos relevantes que se encontram na coleção total de artigos. E o índice de precisão na recuperação se obtém da relação entre os artigos científicos relevantes recuperados e o total de artigos recuperados.

A análise dos resultados realizou-se a partir da análise quantitativa e análise qualitativa. A análise quantitativa realizou-se a partir dos cálculos dos índices de consistência entre a indexação automática e indexação manual e dos índices de exaustividade e precisão na recuperação da informação, os quais subsidiaram o desenvolvimento da análise qualitativa.

A análise qualitativa, por sua vez, foi realizada por análise e interpretação da fundamentação teórica sobre o processo de indexação, a automatização da indexação, os software de indexação presente na literatura da área e avaliação da indexação. A análise dessas informações e todo processo de preparação das fontes, os testes e a indexação definitiva pelo SISA permitiram constatar os aspectos decisivos na atuação de um software de indexação. Além disso, a avaliação mediante a comparação entre a indexação automática e manual ofereceu dados que exigiram uma análise crítica sobre as possíveis motivos que culminaram nesses resultados apresentados na próxima seção.

5 Resultados

A análise da literatura de Ciência da informação sobre indexação e automatização da indexação mostrou que a indexação automática possui metodologias muito diferentes em relação à manual, apesar de possuírem o mesmo objetivo. O método de indexação aplicado na atribuição de descritores na indexação manual é baseado na investigação do significado dos conceitos presentes no artigo, havendo um processo de reflexão sobre os assuntos, o que caracteriza a atividade do processo cognitivo humano na atribuição de significado durante a compreensão. Enquanto na indexação automática do SISA verifica-se um processo diferente, ou seja, processo automático baseado na freqüência de palavras que responde a regras determinadas e isso tem influência no uso da linguagem documentária.

Os resultados obtidos se referem: à análise da indexação com o SISA; à análise da consistência entre indexação manual da BIREME e a indexação automática do SISA e; à avaliação da exaustividade e precisão recuperação da informação, apresentados à seguir.

5.1 Análise do processo de uso do SISA

O processo de uso do software SISA exigiu a estruturação das fontes: lista de descritores, lista de palavras vazias e os artigos científicos, o que possibilitou verificar com detalhes, a importância de uso adequado para obter êxito em seu funcionamento.

Na estruturação da lista de descritores constatou-se a importância do recurso remissivo por meio da palavra USE que proporciona maiores possibilidades de um assunto ser atribuído, mesmo que esse assunto esteja representado em diversas formas, como veremos na análise da consistência.

A utilização da lista de palavras vazias e a interferência que esta tem sobre a atribuição de termos de indexação, especialmente na atribuição de termos compostos, foram percebidas durante os testes que revelaram a necessidade de verificar se preposições presentes em termos compostos como em *diagnóstico por imagem* se encontravam na lista de palavras vazias, pois situações dessa natureza impediriam ao software atribuir um termo composto.

A estruturação dos artigos científicos mediante marcadores de título, resumo e texto demonstra claramente o método de seleção de termos pelo SISA, que estabelece a combinação da freqüência dos termos nessas partes do artigo para propor os termos de indexação.

Portanto, a aplicação do SISA foi importante, não apenas para obter os termos de indexação, mas também para verificar os fatores que existem por trás do funcionamento do software e compreender as conseqüências que estes provocam na consistência da indexação e na exaustividade e precisão da recuperação da informação.

5.2 Análise da consistência de indexação

Os índices de consistência de atribuição de termos entre as duas formas de indexação variam de 0% a 75% com um índice médio de 23,25%.

A partir da análise comparativa dos descritores que a BIREME e o SISA atribuíram, verificamos a necessidade de analisar as razões que impediram o SISA de atribuir alguns descritores determinados pela BIREME:

a) Os artigos científicos apresentam termos diferentes dos que estão na linguagem documentária DeCS. Como no exemplo: *adenoma pleomórfico* utilizado no artigo e *adenoma pleomorfo* no DeCs;
b) Os artigos científicos apresentam termos no plural enquanto no DeCS se encontram no singular. E vice-versa. Exemplos: *osteomas* encontrado no artigo e *osteoma* no DeCS. E o termo *laser* no artigo e somente *lasers* no DeCS;
c) Os artigos científicos apresentam termos simples enquanto no DeCS se encontram apenas termos compostos. Exemplo: O termo *erosão* é utilizado isoladamente com mais freqüência no artigo e no DeCS é encontrado somente o termo *erosão dentária*;
d) Os artigos científicos apresentam termos compostos com pequenas variações em relação aos termos compostos do DeCS. Exemplo: No artigo encontra-se o termo *câncer de laringe* e no DeCS encontra-se a remissiva *câncer da laringe use neoplasias laríngeas*, ou seja, não será possível atribuir *neoplasias laríngeas* por conta da diferença na preposição da para de;
e) Os termos dos artigos científicos que são representativos do conteúdo temático se encontram com freqüência relativamente alta apenas em uma parte da estrutura do artigo. Como no caso do termo *resinas compostas* que se encontra apenas no "texto" do artigo.

A linguagem utilizada pelos autores dos artigos científicos não coincide com os termos que se encontram no DeCS, verificado em casos como: *câncer da região de cabeça e pescoço* ao invés de *câncer da cabeça e pescoço*, *teste da vitalidade pulpar* ao invés de *teste da polpa dentária*, *síndrome da ardência bucal* ao invés de *síndrome da boca ardente*. Além de situações em que pequenas variações impedem a atribuição de termos relevantes: *gengivoestomatite herpética* no artigo e *gengivostomatite herpética use estomatite herpética* no DeCS; *imuno-histoquímica* no artigo e somente *imunohistoquímica use imunoistoquímica* no DeCS. Houve também situações em que o uso siglas ao invés de termos compostos impediu a atribuição de descritores, como em: *cpod* no artigo ao invés de *índice cpod* que remeteria a *índice cpo* do DeCS.

Verificou-se que os principais fatores dificultadores da atribuição de descritores mais próximos dos determinados pela BIREME estão relacionadas às variações dos termos utilizados pelos autores dos artigos científicos em comparação aos que se encontram no DeCS. Constatou-se ainda, que o SISA atribuiu muitos termos simples e gerais por conta desses termos ocorrerem com maior freqüência no artigo do que os termos compostos. Muitas vezes os termos utilizados nos artigos não se encontravam no DeCS nem mesmo com remissiva, o que agrava mais a situação por ocorrer omissão de termos. Além disso, as variações de número e também

de gênero dos termos foram um dos principais fatores que impediram uma consistência maior entre os dois tipos de indexação. Esses resultados mostraram a necessidade de adequação da linguagem documentária alfabética principalmente referente à estrutura de relações entre os termos, considerando-se que as relações de significado entre os termos devem ser mantidas durante a indexação por um sistema automático como um software de indexação para garantir qualidade no processo de recuperação da informação.

Constatou-se que os fatores apresentados na análise da consistência foram influenciadores na avaliação da exaustividade e precisão da recuperação da informação. A influência esteve relacionada ao fato de que a impossibilidade de atribuição de descritores pelo SISA tornou inviável a recuperação de alguns artigos científicos e, conseqüentemente interferiu nos índices de exaustividade e precisão na recuperação das informações.

5.3 Análise da recuperação da informação

Os resultados das buscas realizadas nas bases de dados BDSISA e BDBIREME, por meio dos cálculos de exaustividade e precisão na recuperação que são apresentadas no quadro abaixo, permitiram analisar os fatores envolvidos na indexação e que interferiram nos índices.

1. Base de dados BDSISA (Indexação A)	**Base de dados BDBIREME (Indexação B)**	**2. Base de dados BDSISA (Indexação A)**	**Base de dados BDBIREME (Indexação B)**
1ª Busca: Neoplasias das glândulas salivares Recuperação: 0 Exaustividade = 0/3 = 0% Precisão = 0/0= 0%	1ª Busca: Neoplasias das glândulas salivares Recuperação: Artigo 65 Exaustividade=1/3= 0,33= 33% Precisão = 1/1= 1 = 100%	2ª Busca: Neoplasias bucais. Recuperação: Artigos 10, 34 e 50. Exaustividade = 1/3= 0,33 = 33% Precisão = 1/3 = 0,33= 33%	2ª Busca: Neoplasias bucais. Recuperação: Artigos 10, 16, 25, 34, 46, 50, 62, 71. Exaustividade = 3/3=1= 100% Precisão = 3/8 =0,375 =37,5%
3ª Busca: Cistos odontogênicos. Recuperação: Artigos 98. Exaustividade=1/2=0,5= 50% Precisão = 1/1 = 1 = 100%	3ª Busca: Cistos odontogênicos. Recuperação: Artigos 91 e 98. Exaustividade =1/2=0,5= 50% Precisão = 1/2 = 0,5 = 50%	4ª Busca: Osteoma. Recuperação: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	4ª Busca: Osteoma. Recuperação: Artigo 5. Exaustividade = 1/1=1= 100% Precisão = 1/1=1= 100%
...	...	...	...
49ª Busca: Odontoblastos AND Fibroblastos. Recuperação: Artigo 9 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%	49ª Busca: Odontoblastos AND Fibroblastos. Recuperação: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	50ª Busca: Oncogenes. Recuperação: 0 Exaustividade = 0/2= 0% Precisão = 0/0= 0%	50ª Busca: Oncogenes. Recuperação: Artigos 10 e 20. Exaustividade = 1/2= 0,5= 50% Precisão = 1/2= 0,5 = 50%

Quadro 7: Cálculos dos índices de exaustividade e precisão na recuperação

Os índices de exaustividade na recuperação da BDSISA oscilaram de 0 a 100% com o índice médio de 35, 72%, enquanto na BDBIREME houve a mesma oscilação e índice médio de 77, 04%. Os índices de precisão na recuperação também oscilaram de 0 a 100% nas duas bases de dados, mas com diferença no índice médio, que na BDSISA foi de 40, 92% e na BDBIREME foi de 78, 69%.

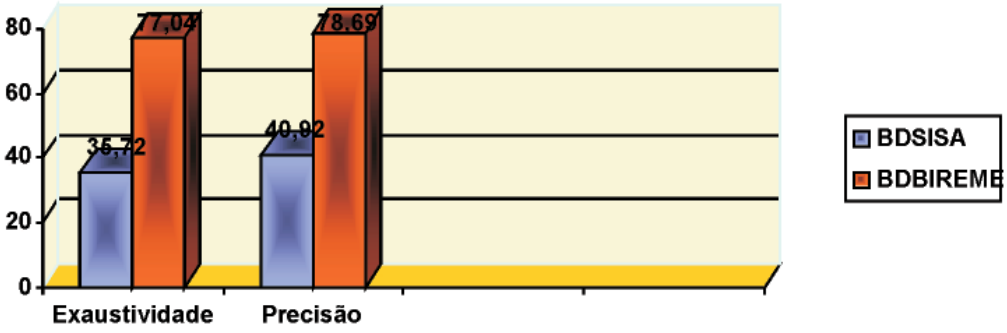

Gráfico 1: Exaustividade e precisão na recuperação nas bases de dados BDSISA e BDBIREME.

Os resultados obtidos na análise de exaustividade e precisão na recuperação comprovaram os fatores que interferiram na análise da consistência e merecem destaque para atuação efetiva de softwares de indexação. Entre os fatores destacamos àqueles que impediram a recuperação de artigos científicos relevantes:

a) A dificuldade do SISA atribuir termos compostos: o SISA atribui mais termos simples ao invés de termos compostos e a estratégia de busca por termo composto torna inviável a recuperação de artigos científicos. Exemplos: O SISA atribuiu os termos *neoplasias* e *glândulas salivares*, enquanto a estratégia de busca se constituiu do termo n*eoplasias das glândulas salivares*. *Neoplasias* e *boca* atribuído pelo SISA e a estratégia de busca *neoplasias bucais*. Os termos *materiais para moldagem odontológica, assistência odontológica para pessoas portadoras de deficiência, cisto odontogênico calcificante* dificilmente são atribuídos pelo SISA, ao contrário da indexação manual que não apresenta dificuldade.

b) A variação dos termos do artigo científico em relação a linguagem documentária DeCS impossibilitou a atribuição pelo SISA de termos como *obturação retrógrada*, pois o termo *retrobturação* é o utilizado no artigo científico. Exemplos: o artigo apresenta o termo *microinfiltração* e, não é atribuído pelo SISA porque esse termo não existe na linguagem documentária e, portanto a estratégia de busca com o termo *infiltração* não pôde recuperar esse artigo. No caso de um artigo com o termo *materiais restauradores* o SISA não atribuiu nenhum termo, ao contrário da BIREME que por um processo de compreensão atribui o termo *materiais dentários,* pelo qual será realizada a recuperação do artigo.

c) Outro aspecto verificado foi o uso no artigo científico apenas de termos relacionados com o termo relevante, como em *oclusão dentária bilateral* ao invés de *oclusão dentária balanceada* da linguagem documentária DeCS. Exemplos: O uso no artigo científico de termos como *pacientes geriátricos* e *idosos* em vez de *odontologia geriátrica*, *clareadores* e *clareamento* ao invés de *clareamento de dente*.

d) O SISA não atribuiu o termo que se encontra apenas em uma parte da estrutura do artigo científico. Exemplo: O termo *palato* se encontra apenas na estrutura do "texto" e, portanto não é atribuído como descritor. O mesmo ocorre com termos como *cisto dentígero* e *fluorose dentária*.

Considerando esses fatores, verifica-se a possibilidade de aperfeiçoamento da linguagem documentária através do acréscimo de termos não preferidos remetendo aos preferidos tendo em vista todas as variações que possam ocorrer nos termos, a partir de uma análise detalhada da terminologia e da literatura da área em que será aplicado o software de indexação. Outro aspecto para aperfeiçoar a linguagem documentária é o estudo dos termos compostos, porque verificamos que existe maior probabilidade

de coincidência entre os termos da linguagem documentária e do artigo científico no caso de termos simples. Constatamos que na indexação automática é necessário compreender e considerar principalmente o método pelo qual o software realiza a análise de conteúdo temático dos documentos para que o uso da linguagem documentária possa se realizar de modo que contemple os critérios do software de indexação e permita indexação de qualidade.

Associado às modificações na linguagem documentária é interessante ao SISA incorporar métodos lingüísticos fundamentados na análise de nível morfológico e sintático que permitem formalizar e distinguir as variações que ocorrem nos termos. Cabe lembrar que as regras estabelecidas pelos algoritmos não contemplam todas as situações que ocorrem em uma língua, pois esta é caracterizada justamente pela flexibilidade e riqueza de significados que estão além da estrutura e posição das palavras, entretanto, esses métodos com fundamentação lingüística poderão minimizar os problemas envolvidos na análise do conteúdo temático dos documentos.

6 Considerações finais

Em resumo, podemos dizer que as principais conclusões foram as seguintes:

1 Com relação ao uso do SISA, verificou-se a importância da estruturação adequada das fontes (artigo científico, lista de descritores e lista de palavras vazias) para adequado funcionamento do SISA.

2 A falta de uma flexibilidade na indexação automática impediu a atribuição de termos relevantes para indexação. Isso é explicado pela incompatibilidade existente entre os termos do artigo científico identificados pelo software SISA e os termos da linguagem documentária DeCS que impediu a determinação de muitos termos atribuídos pela BIREME. Além disso, o SISA atribuiu muitos termos simples pela dificuldade em atribuir termos compostos.

3 Os fatores que dificultaram a obtenção de um nível mais alto no índice de consistência da indexação foram influenciadores dos índices de exaustividade e precisão da recuperação da informação. Essa influência está relacionada à impossibilidade de atribuição de descritores pelo SISA que tornou inviável a recuperação de alguns artigos científicos.

4 Há necessidade de estudos em torno da adequação da linguagem documentária ao uso do software SISA a partir da incorporação e avaliação de métodos lingüísticos em nível de análise morfológica e sintática.

5 Há necessidade da convergência do conhecimento de diversas áreas, entre as quais, a Lingüística computacional, a Ciência da computação, a Estatística, a Ciência da Informação, para o pleno desenvolvimento da área de automatização da indexação.

Portanto, para automatizar a indexação é necessário considerar esse processo inserido no sistema de informação como um todo, ou seja, levar em conta a existência de inúmeras variáveis que interferem em seus resultados. Entre essas variáveis podemos citar a área de conhecimento, a linguagem documentária, os usuários, recursos do sistema de informação principalmente os profissionais, entre tantas outras. Por isso, torna-se imprescindível efetuar um planejamento minucioso buscando prever as variáveis que poderão influir no desenvolvimento e, conseqüentemente no uso e no objetivo do software de indexação. A aplicação de um software de indexação exige do profissional da informação uma postura crítica no monitoramento e na avaliação do software para tornar possível seu aperfeiçoamento, já que sempre haverá a necessidade de aperfeiçoá-lo diante das limitações e dos avanços tecnológicos que oferecem melhorias a serem implementadas de acordo com as necessidades e possibilidades do sistema de informação.

Entretanto, para que esse aperfeiçoamento possa se concretizar, os profissionais da informação precisam compreender melhor como atuam e as possibilidades que os softwares de indexação oferecem e, dessa forma contribuam para desenvolver aplicativos que considerem as reais necessidades de um sistema de informação.

AUTOMATIZED INDEXING OF SCIENTIFIC PERIODICAL PAPERS: ANALYSIS OF THE APPLYING OF SISA SOFTWARE USING DeCS TERMINOLOGY IN DENTISTRY AREA

Abstract The indexing automation has been discussed by researches in the area of Information Science however the discussions have not been so clear on the use of indexing software. Thus, it is necessary to know the indexing software, as well as its application in the analysis of documentary contents. To do so, it is proposed, here, to investigate both the consistency of indexing and the exhaustiveness and precision of the information retrieval, by means of comparative analysis between SISA (Sistema de Indización Semi-Automatico) automatic index and BIREME (Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde) manual indexing. The aim of this paper is to contribute to the theoretical development of the indexing automation and the improvement of SISA. Thus, SISA application and evaluation was used based on the calculation of the consistency indexes between the two types of indexing, and the calculation of the exhaustiveness and precision indexes in information retrieval, by means of searching into BDSISA and BIREME databases, composed by descriptors taken from SISA and manual indexing respectively. The differences among the terms used in scientific papers comparing to the DeCS ones were the main difficult factor to achieve higher consistency indexes in the indexing. These differences influenced the exhaustiveness and precision indexes in the information retrieval, showing that it is necessary to improve the documentary language used by SISA software and to incorporate linguistic methods.

Keywords: Automatized Indexing; Indexing Software; Semi-Automatic Indexing System (SISA), Indexing evaluation.

Artigo recebido em 11/02/2009 e aceito para publicação em 30/04/2009

Referências

GIL LEIVA, I. **La automatización de la indización de documentos**. Gijón: Trea, 1999. 221 p.

_____. Sistema para la Indización Semi-Automática (SISA) de Artículos de Revista de Biblioteconomía y Documentación. In: JORNADAS DE TRATAMIENTO Y RECUPERACIÓN DE INFORMACIÓN, 2., 2003, Leganés (Madrid), Anais eletrônico… Leganés (Madrid), 2003. p. 228-232. Disponível em: <http://webs.um.es/isgil> Acesso em: 13 02 2008.

_____. **Manual de indización.** Teoría y práctica. Gijón: Trea, 2008, 429 p. Sumário e prólogo da obra disponível em: webs.um.es/isgil

GUIMARÃES, J. A. C. **Indexação em um contexto de novas tecnologias**. 2000. 10 p. Texto didático.

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

MENDEZ RODRÍGUEZ, E. M., MOREIRO GONZÁLEZ, J. A. Lenguaje natural e indización automatizada. **Ciencias de la Información**, v. 30, n.3, p.11-24, set., 1999.

MOREIRO GONZÁLEZ, J. A. **El contenido de los documentos textuales:** su análisis y representación mediante el lenguaje natural. Gijón (Astúrias): Trea, 2004. 291 p.

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**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. e ampl. Brasília DF: Edição de autor, 2005. 410 p.

_____. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, v. 20, p. 130-6, jul.dez. 1991.

SILVA, M. R.; FUJITA, M. S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, v. 16, n. 2, p. 133-161, maio/ago., 2004. Disponível em: <http://revistas.puc-campinas.edu.br/transinfo/viewarticle.php?id=65>. Acesso em: 25 abr. 2007.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, Brasília, v. 17, n.1, p.43-57, jan./jun., 1988.