Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
ISSN: 1518-2924
adilson.pinto@ufsc.br
Universidade Federal de Santa Catarina
Brasil

de Brito SILVA, Sâmela Rouse; Fernandes CORREA, Renato
Sistemas de Indexação automática por atribuição: uma análise comparativa
Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, vol. 25, 2020, -, pp. 1-25
Universidade Federal de Santa Catarina
Brasil

DOI: https://doi.org/10.5007/1518-2924.2020.e70740

Artigo Original

Encontros Bibli

SISTEMAS DE INDEXAÇÃO AUTOMÁTICA POR ATRIBUIÇÃO: uma análise comparativa

Systems for automatic indexing by assignment: a comparative analysis

Sâmela Rouse de Brito SILVA
Biblioteca Central, Universidade Federal de Alagoas, Maceió, Brasil
ssamela.brito@hotmail.com
https://orcid.org/0000-0002-9694-2187

Renato Fernandes CORREA
Departamento de Ciência da Informação,
Universidade Federal de Pernambuco, Recife, Brasil
renato.correa@ufpe.br
https://orcid.org/0000-0002-9880-8678

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: Analisa comparativamente dois sistemas de indexação automática por atribuição multilíngue: SISA e MAUI. O SISA (Sistema de Indexação Semiautomático) foi desenvolvido na Espanha, sendo inicialmente proposto para a área de Biblioteconomia e Documentação. Trata-se de um sistema especialista que indexa de forma automática seguindo um tesauro e regras predeterminadas de indexação com base na frequência e posição dos termos. O MAUI (*Multi-purpose Automatic Topic Indexing*) é um sistema de origem neozelandesa que apresenta a especificidade de utilização de um tesauro e algoritmo de aprendizagem de máquina para gerar modelo a partir de resultados da indexação intelectual, sendo os termos representados por características estatísticas.
Método: A pesquisa se classifica como exploratória e bibliográfica, onde o método utilizado para construção deste trabalho foi o estudo comparativo baseado na análise de conteúdo das publicações científicas contendo relatos de experiência na aplicação dos sistemas. As etapas da pesquisa consistiram em descrever e comparar as características de cada sistema, levantando informações acerca de como são processados os documentos, como é feita a extração e seleção dos termos que resulta nos descritores propostos por cada sistema, e contextos de aplicação.
Resultado: Como resultados aponta-se as abordagens, as principais operações, os recursos utilizados por cada sistema durante o processamento da indexação automática, bem como os contextos de uso e qualidade alcançada nos resultados.
Conclusões: O trabalho contribui para os estudos na temática indexação automática no aprofundamento da discussão sobre características descritivas e comparativas associadas aos métodos e técnicas implementadas nos sistemas analisados.
PALAVRAS-CHAVE: Indexação Automática. Indexação Automática por Atribuição. Sistema de Indexação Automática. Processamento de Linguagem Natural. Recuperação da Informação.

ABSTRACT

Objective: This work presents a comparative analysis between two multilingual automatic indexing systems that perform term assignment: SISA and MAUI. The SISA (Semi-automatic Indexing System) made in Spain and initially proposed for the area of Librarianship and Documentation, it is a specialist system that automatically indexes following a thesaurus and predetermined rules of indexation which are based on the frequency and position of the terms. The MAUI (Multi-purpose Automatic Topic Indexing) is a system of New Zealand origin that presents the specificity of use of a thesaurus and algorithm of machine learning to generate model through the results of the intellectual indexing, being the terms represented by statistical features.
Methods: The research is exploratory and bibliographical, where the method used to construct this work was the comparative study based on content analysis of the scientific publications containing experience reports of application of that software. The stages of the research consisted of describing and comparing the characteristics of each system, raising information about how the documents are processed, how the systems performs the extraction and selection of the descriptors terms, and the application context.
Results: The results show the approaches, main operations, the resources used by each system during the automatic indexing process, as well as the application context and quality of results.
Conclusions: It hopes to contribute to the studies on the topic of automatic indexing in the deepening discussion about descriptive and comparative categories related to methods and techniques implemented in the systems.
KEYWORDS: Automatic Indexing. Automatic Indexing by Assignment. Automatic Indexing Systems. Natural Language Processing. Information retrieval.

1 INTRODUÇÃO

Com o avanço da ciência e da tecnologia, um volume crescente de publicações é gerado pela comunidade acadêmica em diversos domínios e rapidamente disseminado através da Web.

Neste contexto, visando que os pesquisadores consigam encontrar publicações relevantes para fins de pesquisa, o processo de indexação de tais documentos se constitui como o fator primordial para alcance de qualidade na recuperação da informação (GIL LEIVA, 2009) e consequente sucesso de qualquer mecanismo de busca ou base de dados científica.

O uso da tecnologia no processo de indexação tem sido bastante discutido na área da Ciência da informação. Isto pode ser explicado pela necessidade de se automatizar o processo de indexação diante da rapidez exigida aos profissionais indexadores na análise de cada documento.

Em suas pesquisas, Lancaster (2004) e Moreiro González (2004) acreditam que a indexação automatizada pode contribuir consideravelmente em torno das tarefas do indexador ao maximizar seu tempo de trabalho nas diversas unidades informacionais existentes.

Os sistemas de indexação automática realizam a indexação mediante atividade de análise do texto do documento por um sistema computacional, sem que haja uma interferência humana, isto é, os termos de indexação são definidos apenas pela análise realizada pelo software (LEIVA, 1999). Entretanto, para que se alcance uma melhor qualidade no processo de indexação, acredita-se que seja necessária a intervenção do indexador na avaliação e posterior validação dos termos propostos por tais sistemas para os documentos (NARUKAWA; GIL LEIVA; FUJITA, 2009).

A indexação automática pode ser por extração ou por atribuição. A indexação automática por extração, denominada em (LANCASTER, 2004) como indexação por extração automática, consiste em extrair do texto do documento termos e depois ponderar e selecionar os termos mais relevantes. A indexação automática por atribuição é denominada em (LANCASTER, 2004) como indexação por extração automática, e consiste em associar a termos autorizados de um vocabulário controlado um perfil formado por um conjunto de termos alternativos ou expressões equivalentes que ocorrem

com frequência nos documentos, os termos autorizados que possuem ocorrências em um documento são ponderados e posteriormente são selecionados os mais relevantes.

Este trabalho tem como ponto de partida a comparação de dois sistemas de indexação automática por atribuição multilíngue: SISA e MAUI. A escolha em pesquisar sistemas de indexação automática por atribuição se deu pela mesma permitir o controle terminológico e uso de tesauro na indexação e recuperação da informação (BANDIM; CORREA, 2019).

O SISA (Sistema de Indexação Semiautomático) foi desenvolvido na Espanha, sendo inicialmente proposto para a área de Biblioteconomia e Documentação. Trata-se de um sistema especialista que indexa automaticamente seguindo um vocabulário controlado e regras predeterminadas de indexação com base na frequência e posição dos termos.

O MAUI (*Multi-purpose Automatic Topic Indexing*) é um sistema multilíngue de origem neozelandesa que faz uso de um tesauro e algoritmo de aprendizagem de máquina para gerar um modelo a partir de resultados da indexação intelectual, sendo os termos representados por características estatísticas.

O objetivo deste trabalho é realizar uma análise comparativa dos sistemas SISA e MAUI, levantando características descritivas e comparativas associadas à operabilidade e especificidades no processamento da indexação automática por atribuição. O procedimento metodológico tem como base a análise de conteúdo de relatos de experiência de aplicações dos sistemas.

A presente pesquisa justifica-se pelo fato de que a indexação automática por atribuição é vista como necessária para que o processo de indexação se torne mais ágil e produtivo para o indexador, além de possibilitar a realização da atividade profissional em múltiplos domínios especializados. Adicionalmente, autores desconhecem algum trabalho na literatura nacional de Ciência da Informação acerca da análise comparativa de sistemas de indexação automática por atribuição. Fato que ressalta a originalidade, pioneirismo e relevância do presente trabalho.

Para atingir os objetivos, foi feito um levantamento e análise de relatos de experiência da aplicação dos dois sistemas registrados em publicações científicas, buscando sintetizar a descrição dos mesmos quanto às características e contextos de aplicação, e posteriormente compará-las.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção trata do referencial teórico acerca do processo de indexação, destacando assim a indexação automática. Na subseção 2.1, discute-se o processo de indexação, conceitos e tipologias. Em seguida, na subseção 2.2, discute-se a indexação automática na perspectiva da organização da informação e no contexto de recuperação da informação.

2.1 Indexação: Conceito e Tipologia

Os profissionais da informação realizam a indexação para representar as informações em documentos visando a posterior recuperação.

De acordo com Cintra (1983) “o processo de indexação consiste na tradução de um documento em termos documentários, isto é, em descritores, cabeçalhos de assunto, termos-chave, que têm por função expressar o conteúdo do documento”.

A organização e recuperação da informação se materializam pela indexação, que por sua vez é realizada com a finalidade de determinar, por meio do conteúdo dos documentos, um conjunto de palavras-chave ou assuntos, facilitando sua armazenagem em bases de dados e atendendo deste modo, as necessidades de recuperação da informação (FUJITA; GIL-LEIVA, 2010).

Para realizar a indexação é necessário que o documento seja minuciosamente analisado pelo indexador e que sejam estudadas as estratégias de busca, as necessidades de informação e o perfil do usuário efetivo e/ou potencial daquela determinada unidade informacional para qual a indexação está direcionada.

Para Lancaster (2004), a indexação envolve duas etapas básicas: análise de assunto e tradução. A análise de assunto ou análise conceitual refere-se ao assunto de que trata o documento. A tradução refere-se à conversão da análise conceitual de um documento em um conjunto de termos de indexação. Porém, outras etapas podem ser encontradas na literatura (GIL LEIVA, 2009).

Lancaster (2004) também classifica a indexação em intelectual, semiautomática e automática. Segundo Narukawa (2011), os três tipos de indexação podem ser distinguidos da seguinte forma:

- Indexação Intelectual: também denominada indexação manual, é caracterizada pela atividade humana durante todo o processo, não utilizando assim de qualquer dispositivo tecnológico. Portanto, o homem é o responsável pela realização de todas as etapas do processo de indexação.
- Indexação Semiautomática: relaciona-se ao processo em que um sistema computacional realiza a atividade de análise de texto do documento e, posteriormente, um indexador humano avalia os termos para indexação indicados pelo sistema, escolhendo assim, quais termos serão efetivamente empregados para representar o documento.
- Indexação Automática: consiste da seleção automática dos termos por um software ou sistema, levando em conta critérios estatísticos da ocorrência dos termos no texto dos documentos.

O tipo de indexação dependerá da instituição e dos recursos disponíveis para tal. Entretanto, independentemente de a indexação ocorrer ou não por meio da aplicação de tecnologia, o indexador sempre buscará a representação e a recuperação da informação contida nos documentos do acervo e em benefício ao usuário.

O estudo da indexação automática deriva da necessidade de aprimorar o processo de recuperação da informação ante aos efeitos do avanço da produção de documentos eletrônicos ou digitais. A aplicação da indexação automática desenvolveu-se como uma alternativa viável na análise e representação da informação diante do crescimento exponencial do volume de documentos.

Em decorrência do crescimento documental, Fujita (2003) afirma que:

> O século XIX foi o período em que a indexação começou a apresentar um aprimoramento de sua execução e ao mesmo tempo ser apreciado pelo público, que sentia necessidade de encontrar uma fórmula para controle da massa documental que crescia em demasia (FUJITA, 2003, p. 140).

A construção de índices, considerados os primeiros instrumentos para armazenagem e recuperação de informações, marca essa época em que o homem passa a preocupar-se em registrar e organizar documentos por meio da representação de conteúdo.

Representar conteúdo dentro de um Sistema de Recuperação da Informação (SRI) é considerado um dos atos mais relevantes de uma unidade de informação, uma vez que

este processo reflete a finalidade ou missão junto ao usuário de permitir a recuperação da informação de forma rápida e prática.

Para Silva e Fujita (2004) o conceito de indexação se deu a partir da construção de índices, apesar de nos dias atuais estar mais diretamente vinculada à conceituação e análise de assuntos. Para as autoras, a necessidade dos usuários de uma recuperação mais precisa e especializada, tornou a indexação um aparato metodológico indispensável para instituições informacionais.

Lancaster (2004) afirma que o ato de indexar transcende a ideia de apenas construir índices, uma vez que o avanço tecnológico global exige uma necessidade de condensação mais abrangente para uma recuperação mais aproximada daquilo que necessita o usuário.

Para muitos autores a indexação é discutida como um mecanismo de representação documentária cujo propósito é unicamente voltado para a recuperação da informação contida em determinado documento, como por exemplo, Borko e Bernier (1978) ao afirmarem que a indexação é o processo de analisar o conteúdo informacional de registros informacionais.

Diante dessas definições pode-se inferir a importância da indexação e do seu valor dentro de uma unidade informacional, bem como o papel imprescindível atribuído ao profissional da informação no que tange às atividades de indexação visando a representação e recuperação da informação.

2.2 Indexação Automática: A tecnologia aplicada à representação da informação

Na era da Internet, a partir do momento em que o pesquisador publica os resultados de sua pesquisa, é gerado um produto da ciência cujo objetivo é a ampla disseminação e disponibilização aos pares de forma instantânea.

À medida que o aumento das publicações científicas possibilita novas descobertas e desenvolve a ciência, eis que surge a preocupação referente aos processos de organização dessas publicações, destacando assim, a disseminação e recuperação dessas informações nas bases de dados científicas.

De acordo com Fujita (2003), a organização da informação compreende as atividades e operações de tratamento da informação que abrange conhecimento teórico e

metodológico tanto para o tratamento descritivo, quanto para o tratamento temático de conteúdo das informações. Assim, a recuperação consiste na localização de documentos inseridos em bases de dados e requer técnicas e métodos da organização da informação a fim de satisfazer às necessidades de busca.

Ao passo em que a evolução da tecnologia desenvolve a produção de informações na comunidade científica, ela também possibilita o uso de instrumentos e ferramentas capazes de colaborar nos processos de representação e posterior recuperação das informações. Segundo Robredo (2005), para disponibilizar o acesso rápido às bases de dados científicas, o suporte do computador é de suma importância no processamento de dados e informações.

Na indexação, o uso da tecnologia é denominado por Lancaster (2004) como indexação automática. O autor a define como “um processo que ocorre quando o computador é utilizado para substituir a indexação manual realizada por um indexador”. Para Hjørland (2008), “a indexação automática é realizada por meio de procedimentos algorítmicos”. Segundo Araújo Júnior (2007) a indexação automática é “qualquer procedimento que permita identificar e selecionar os termos que representam o assunto dos documentos sem a intervenção direta do homem”.

Neste tipo de indexação a tecnologia é empregada e o processo acontece automaticamente em diversos ambientes informacionais que fazem uso de mecanismos de busca. A indexação automática é amplamente adotada pelos sistemas de recuperação de informação de bibliotecas e centros de documentação na construção de índices para busca, mas em poucos casos, sua aplicação é feita de forma transparente para o profissional da informação. Os principais motivos da aplicação não consciente de tecnologia no processo de indexação são a falta de conhecimento da existência e do uso de sistemas de indexação automática, e a resistência por dúvidas de como aplicá-la de forma a obter eficácia na representação da informação.

Para Lancaster (2004) a indexação automática pode ser por extração ou por atribuição. O processo se dá por extração quando palavras ou expressões presentes num documento são selecionas para representar seu conteúdo; e por atribuição quando envolve a atribuição de termos a um documento de uma fonte que pode não ser o próprio documento (LANCASTER, 2004). Esta fonte é um vocabulário controlado da área do conhecimento de que se pretende realizar o processo de indexação. O vocabulário controlado consiste numa lista de termos autorizados que forma uma linguagem de indexação, podendo ser um tesauro ou até mesmo uma lista alfabética.

A diferenciação entre a indexação automática por extração e a indexação automática por atribuição ocorre dado que na primeira são extraídos do texto os termos pertinentes à representação do conteúdo, enquanto na segunda os termos extraídos são traduzidos em termos autorizados de um vocabulário controlado, forçando um controle terminológico dos termos que irão representar o conteúdo temático dos documentos.

Um grande diferencial da indexação automática por atribuição em relação à indexação automática por extração utilizada na maioria dos atuais sistemas de recuperação de informação é que a primeira representa o conteúdo do documento por meio de termos autorizados de um vocabulário controlado, e assim, permite uma melhor eficácia na recuperação da informação, já a segunda é utilizada para construir índices de palavras isoladas, não contemplando termos compostos de uma linguagem de especialidade como uma unidade do discurso.

Narukawa, Gil-Leiva e Fujita (2009) discorrerem que o ideal seria que a tecnologia fosse empregada aliada à intervenção humana para avaliar e validar os termos descritores. Os autores apontam que o tempo que é destinado ao processo de indexação intelectual pode ser facilmente agilizado por um sistema automatizado, e, portanto, a tecnologia deve ser empregada contribuindo assim no processo de indexação, apontando deste modo, a indexação semiautomática como a melhor alternativa.

De acordo com Lancaster (2004):

> O software promete aos profissionais da informação uma contribuição no sentido de diminuir o tempo dedicado ao trabalho atribuído ao processo de indexar, tendo em vista que a indexação é um processo que necessita tempo e técnicas para ser considerada de boa qualidade (LANCASTER, 2004).

Assim, a indexação automática é realizada por meio de programas de computador que podem contribuir para agilizar o processo de indexação, economizando tempo gasto em tal atividade. A automatização da indexação serve como ferramenta de auxílio na prática do profissional da informação, contribuindo para que a indexação possa ser realizada de maneira otimizada por meio da tecnologia empregada em sistemas que permeiam as tarefas do indexador. Agilizar o processo intelectual realizado pelos profissionais da área deve ser o principal objetivo das pesquisas sobre a indexação automática.

O estudo acerca da automatização da indexação teve início na segunda metade do século XX. Remonta aos últimos anos da década de 50, quando Luhn apresentou o índice KWIC (*Key Word In Context*), em que as palavras do título que servem de entradas no índice são identificadas automaticamente por meio da eliminação das palavras que não tem significado, por comparação com uma lista de palavras vazias de significado, estabelecida previamente (ROBREDO, 1991). Seu desenvolvimento se deu basicamente em momentos históricos e permeiam fatores linguísticos, estatísticos e híbridos.

No artigo de Corrêa e Lapa (2014), "A indexação automática no Brasil no âmbito da Ciência da Informação (1973-2012): indicadores bibliométricos", os autores apresentam um histórico das pesquisas brasileiras sobre indexação automática, explicam como se deu o seu desenvolvimento e apontam tecnologias empregadas nos sistemas desenvolvidos.

Tanto o artigo de Corrêa e Lapa (2014), como o de Bandim e Corrêa (2018), mencionam os principais sistemas de indexação automática utilizados no Brasil. No Quadro 1, apresenta-se uma síntese dos sistemas de indexação automática citados nos respectivos trabalhos e que possuem relevância para área.

Quadro 1 – Sistemas de indexação automática utilizados no Brasil

SISTEMAS	ORIGEM	DESCRIÇÃO	TIPO DE INDEXAÇÃO AUTOMÁTICA
OGMA	Universidade Federal de Minas Gerais, Brasil, 2008.	Desenvolvido por Luiz Cláudio Gomes Maia durante o doutorado para analisar textos em português por meio da extração de sintagmas nominais e cálculo do peso desses na indexação dos documentos.	Extração
SISA	Universidade de Múrcia, Espanha, 1999-2008.	Desenvolvido por Gil Leiva, extensível para análise de textos em diferentes idiomas e áreas do conhecimento, dado um vocabulário controlado e lista de palavras vazias.	Atribuição
AUTOMINDEX	Universidade de Brasília, Brasil, 1991.	Desenvolvido por Jaime Robredo, é um software que para propor os termos de indexação, confronta as palavras do texto com as de dois antidicionários concomitantes de palavras vazias: um com palavras invariáveis e outro com raízes de palavras não significativas.	Extração

Fonte: Dados da pesquisa.

O surgimento destes sistemas é de fundamental importância para a área. Por meio desses sistemas, os profissionais da informação têm a possibilidade de realizar as atividades de indexação. Adicionalmente, a partir da utilização dos mesmos é possível aprimorá-los e adequá-los para atender às demandas informacionais das unidades de informação.

Percebe-se por meio do Quadro 1, que somente um dos sistemas realiza indexação automática por atribuição, e por isso será analisado neste trabalho.

Neste artigo, visando analisar comparativamente sistemas que realizam a indexação automática por atribuição, serão analisados o software SISA e o software MAUI, de acordo com o percurso metodológico descrito a seguir.

3 METODOLOGIA

Esta pesquisa se classifica como exploratória e bibliográfica, pautada em estudo comparativo dos sistemas SISA e MAUI por meio de coleta e análise de dados descritivos via análise de conteúdo das publicações científicas contendo relatos de experiência da aplicação dos mesmos.

A pesquisa exploratória busca “estudar, explorar o problema a fim de torná-lo explícito e possibilitar a criação de hipóteses” (CAJUEIRO, 2015, p.16). A pesquisa exploratória é, normalmente, o passo inicial no processo de pesquisa pela experiência e um auxílio que traz a formulação de hipóteses significativas para posteriores pesquisas (CERVO; BERVIAN; SILVA, 2007).

Como ponto de partida, foi realizada uma revisão da literatura pertinente à temática de indexação automática por atribuição. Posteriormente foi realizada uma pesquisa bibliográfica por relatos de aplicação dos sistemas SISA e MAUI.

A pesquisa bibliográfica consistiu da busca e análise de trabalhos na literatura nacional e internacional da área de Ciência da Informação sobre indexação automática por atribuição de documentos escritos em português. Sendo posteriormente filtrados e selecionados os trabalhos que aplicassem o software SISA. Não foi encontrada menção ao software MAUI na literatura de Ciência da Informação, assim, os trabalhos sobre tal sistema foram buscados na literatura de Ciência da Computação.

A partir da pesquisa bibliográfica foram selecionados artigos da área da Ciência da Informação em que o SISA foi aplicado em documentos em língua portuguesa e foram reportadas métricas de qualidade na indexação automática, a saber: (BANDIM; CORRÊA, 2018); (BANDIM; CORRÊA, 2019); (GIL LEIVA, 2017) (NARUKAWA, 2011); e (NARUKAWA, GIL LEIVA e FUJITA, 2009). Esses trabalhos foram então selecionados para análise, sendo o trabalho de GIL LEIVA (2017) uma publicação internacional, e os outros sendo publicações nacionais.

Com relação ao software MAUI, não foi encontrado nenhuma publicação sobre o mesmo na literatura de Ciência da Informação. Acredita-se que o motivo seja o fato de o software ser mais conhecido na área de Ciência da Computação no âmbito internacional. Assim, foram selecionados alguns trabalhos internacionais na área da Computação em que houve aplicação do MAUI a textos científicos e participação do desenvolvedor do software como autor, tais como: (KIM *et al*., 2012); (MEDELYAN; FRANK; WITTEN, 2009) e (MEDELYAN, 2009).

Justifica-se a escolha dos trabalhos selecionados por abordarem a descrição dos respectivos sistemas de maneira elucidativa para os fins deste estudo.

Posteriormente foi realizada a análise de conteúdo das publicações com o intuito de descrever os sistemas e contextos de uso ou aplicação.

Para a coleta e análise dos dados foi empregada a análise de conteúdo, que segundo Bardin (1991) é definida por:

> Um conjunto de técnicas de análise das comunicações, visando obter, por procedimentos, sistemáticos e objetivos de descrição do conteúdo das mensagens, indicadores (quantitativos ou não) que permitam a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) destas mensagens (BARDIN 1991).

Durante a análise de conteúdo dos relatos de aplicação dos sistemas, foram sendo identificadas nos textos categorias descritivas dos sistemas, perpassando pelos requisitos de entrada, operacionalidade e funcionamento dos mesmos, além da identificação de contextos de aplicação.

Foram consideradas as seguintes categorias na descrição dos sistemas: arquivos utilizados no processamento da indexação; formato dos arquivos de entrada; utilização de linguagem documentária; flexibilidade quanto ao idioma e domínio; etapas de processamento; operações sobre o texto; ponderação dos termos; e abordagem adotada

por cada sistema na indexação automática por atribuição. Visando o melhor entendimento do leitor, as categorias citadas serão descritas na seção 4.3 juntamente com a descrição dos valores das mesmas para os sistemas analisados.

Por fim as descrições dos sistemas foram compiladas e analisadas comparativamente com base nas categorias selecionadas e seus respectivos valores.

4 ANÁLISE DE RESULTADOS

Nesta seção, inicialmente são apresentadas e discutidas as descrições dos sistemas SISA e MAUI, criadas a partir da síntese dos relatos de experiência da aplicação dos mesmos. Em seguida, apresenta-se a análise comparativa acerca das características dos sistemas.

4.1 SISA

No artigo (NARUKAWA, LEIVA e FUJITA, 2009), os autores apontam que o software SISA (acrônimo do espanhol *Sistema de Indización Semiautomático*) foi desenvolvido no período de 1999 a 2008 pelo professor e pesquisador Isidoro Gil Leiva, da Universidade de Múrcia, na Espanha.

Originalmente o SISA foi proposto como um sistema de indexação semiautomático aplicado à área da Biblioteconomia e Documentação, porém pode ser aplicado a qualquer área do conhecimento, desde que sua configuração contemple as exigências do software (NARUKAWA, 2011, p.106).

A seguir são descritos os requisitos de entrada de dados do SISA:

- Lista alfabética de termos e descritores: arquivo texto contendo os termos autorizados e termos alternativos, sendo para cada termo alternativo indicado o respectivo termo autorizado;
- Lista alfabética de termos e respectivos termos gerais: arquivo opcional contendo o termo autorizado e a relação de termo geral;

- Lista de palavras vazias no idioma do texto dos documentos, para fins de eliminação das palavras consideradas vazias (*stopwords*), como conectivos e artigos;
- Marcação das partes estruturais constituintes do documento a ser indexado;
- Todos os arquivos de entrada, incluindo os textos a serem indexados, devem estar no formato txt.

Assim, os arquivos utilizados no processamento do SISA são: o texto completo (com título, resumo e texto identificados); uma lista de palavras vazias (*stoplist*); e uma linguagem documentária como, por exemplo, um tesauro da área do conhecimento dos documentos.

Durante a análise de texto do documento, o software adota uma metodologia por meio da comparação entre termos do documento (posicionados no título, resumo ou texto) e termos de uma linguagem documentária. Assim, o SISA identifica termos de um vocabulário controlado no texto completo de um documento, identificando sua ocorrência nos campos estruturais: título, resumo e no texto.

Para que determinado termo do vocabulário controlado seja proposto como termo de indexação, é necessário que ele satisfaça alguma das regras de ponderação envolvendo critérios preestabelecidos de frequência e posição no texto (BANDIM; CORRÊA, 2018).

Além dos termos de indexação sugeridos, o sistema lista também termos candidatos à indexação, que são os termos que não são palavras vazias e não estão no vocabulário controlado, mas que atendem aos critérios de frequência e posição dos termos sugeridos. De acordo com Gil-Leiva (2017), termos candidatos são:

> Termos não incluídos no vocabulário controlado, nem na lista de stopwords, e que cumprem requisitos como aparecendo um número mínimo de vezes e em diferentes parágrafos. Assim, a indexação do SISA não depende unicamente da presença ou ausência de um termo do vocabulário, e também é possível realizar um feedback automático do vocabulário (GIL LEIVA, 2017).

Após o sistema realizar a indexação automática para um documento, o indexador pode escolher descritores entre os termos sugeridos e termos candidatos, mantendo assim a característica de sistema semiautomático de indexação.

De acordo com Narukawa, Leiva e Fujita (2009), a metodologia que envolve o processo de indexação pelo SISA se dá em três etapas:

1. Pré-processamento: nesta fase as partes constituintes do documento devem ser marcadas pelo indexador para sua identificação por meio dos símbolos: Título que deve ser delimitado por #CTI# e #FTI#; Resumo que deve ser delimitado por #CR# e #FR#; e Texto que deve ser delimitado por #CTE# e #FTE#. É nessa fase também que o software realiza a eliminação das palavras vazias e horizontalização das frases, desta forma, o total de palavras nas fontes título, resumo e texto são computadas;
2. Análise de conteúdo: nesta fase ocorre a etapa de análise do texto, onde um algoritmo busca a extração de termos da linguagem documentária que coincidem com termos das fontes;
3. Valoração e ponderação: esta última etapa consiste na aplicação de critérios de avaliação dos termos para que o sistema possa selecionar os termos de indexação que representarão o conteúdo do documento. Na seleção e proposição dos termos para indexação são aplicados os seguintes critérios para que um termo seja considerado como termo de indexação: um termo autorizado aparece no título e no resumo, ou um termo autorizado aparece no título e no texto, ou um termo autorizado aparece no resumo e no texto do documento. No entanto, os termos semivazios são apresentados como candidatos à indexação se, a palavra semivazia aparece no título, resumo e texto, ou se aparece no texto dez vezes ou mais, além de aparecer em oito parágrafos diferentes ou mais. Assim, o sistema irá selecionar e propor os termos de indexação de acordo com regras heurísticas envolvendo a frequência e posição dos termos em relação ao documento (GIL LEIVA, 2017).

Para concluir a indexação semiautomática, é necessário que haja a intervenção humana para analisar e decidir os termos de indexação propostos pelo sistema. Por ser um sistema flexível, é possível acrescentar ou suprimir termos, isso permite que o indexador possa tomar a decisão considerando as particularidades de um sistema de informação (NARUKAWA, LEIVA e FUJITA, 2009).

4.2 MAUI

O termo MAUI é acrônimo do inglês *Multi-purpose Automatic Topic Indexing*, que pode ser traduzido como indexação automática por tópicos de propósito geral.

O software foi desenvolvido por Alyona Medelyan como parte de seu projeto de doutorado, sob a supervisão de Ian H. Witten e Eibe Frank, no Departamento de Ciência da Computação da Universidade de Waikato, Nova Zelândia em 2009.

O MAUI baseia-se no algoritmo de extração de palavras-chave (do inglês *keyphrase extraction*) do software KEA, e fornece funcionalidades adicionais como atribuição de termos de vocabulário controlado, e opcionalmente a atribuição de tópicos a documentos baseados em termos da Wikipedia usando o software Wikipedia Miner. Esses recursos adicionais possibilitam ao MAUI identificar termos de indexação com mais precisão segundo Medelyan, Frank e Witten (2009).

O MAUI permite executar as seguintes tarefas (MEDELYAN, 2009): atribuição de termos de um vocabulário controlado ou tesauro; indexação de assunto; indexação de tópico com termos da Wikipedia; extração de palavras-chave; extração de terminologia; marcação automática; extração de terminologia e indexação de tópico semiautomática; e extração de palavras-chave tendo como referência um vocabulário controlado.

O MAUI extrai automaticamente os principais termos de documentos de texto, e a depender da tarefa, os termos podem corresponder a tópicos, *tags*, palavras-chave, frases-chave, termos de vocabulário, descritores, termos de índice ou títulos de artigos da Wikipédia.

O MAUI pode ser usado como uma ferramenta de sugestão de marca ou *tag* que fornece aos usuários *tags* descrevendo os principais tópicos de documentos recentemente adicionados. As *tags* podem ser corrigidas ou aprimoradas por marcas pessoais, se necessário, o que pode melhorar a consistência de uma folksonomia, por exemplo, sem comprometer a sua flexibilidade.

A metodologia e etapas de processamento segundo Medelyan (2009) são:

1. Geração de tópicos candidatos – extração de termos candidatos a termos de indexação;
2. Cálculo de características – cálculo de características para os termos candidatos;
3. Construção do modelo de indexação – treinamento do modelo de indexação levando em conta os termos atribuídos pelos indexadores a cada documento do conjunto de treinamento;
4. Aplicação do modelo aprendido para selecionar tópicos em outros documentos – aplicação do modelo de indexação treinado para propor termos de indexação a outros documentos.

O software apresenta a especificidade da utilização de um algoritmo de aprendizagem de máquina para gerar um modelo de seleção de termos de indexação como base no comportamento humano de indexação, e esta característica o aproxima da indexação manual.

Para representar o conteúdo dos documentos, o MAUI utiliza métodos estatísticos e linguísticos para a extração e ponderação de termos.

A ponderação de termos leva em consideração o cálculo de características baseadas na frequência de ocorrência do termo (do inglês *frequency features*), na posição de ocorrência do termo (do inglês *occurence features*), o comprimento do termo em palavras (do inglês *length feature*), na probabilidade de um termo ser palavra-chave no domínio ou corpus (do inglês *keyphraseness feature*), e baseada em relações semânticas dos termos em um tesauro (do inglês *thesaurus feature*). Para indexação automática de documentos utilizando termos da Wikipedia, o software também permite o uso características baseadas na Wikipedia (do inglês *Wikipedia features*).

Medelyan, Frank e Witten (2009) descrevem como as características são utilizadas na ponderação dos termos:

- Frequência: São computadas as medidas TFIDF, TF e IDF para cada termo candidato. A medida TFIDF combina a frequência de uma frase em um documento particular (TF) com sua frequência inversa de ocorrência no conjunto de treinamento (IDF). Esta pontuação é alta para termos raros e que aparecem com frequência em um documento, sendo mais propensos a serem significativos.
- Ocorrência: Leva em consideração a posição de ocorrência do termo candidato no documento. A posição de ocorrência é computada como a distância relativa da ocorrência do termo desde o início do documento, normalizada pelo tamanho do documento em palavras. São computadas a primeira (FIRST), última (LAST) e intervalo (SPREAD) de ocorrência de cada termo candidato, sendo este último a diferença entre as duas primeiras. Termos candidatos com valores muito baixos ou muito altos para estas características provavelmente são *tags* ou descritores, porque eles aparecem na abertura do documento, em partes tais como título, resumo e introdução, ou nas seções finais do documento, como conclusão e referências;

- Comprimento: O comprimento do termo candidato ou frase (LENGTH) é medido em palavras. De um modo geral, quanto mais longa a frase, mais específica ela é. Geralmente, indexadores profissionais preferem expressões do que palavras isoladas como palavras-chave. O treinamento do modelo de aprendizagem de máquina captura e quantifica a preferência de especificidade dos termos em um determinado corpus;
- Probabilidade de ser palavra-chave: Quantifica a frequência com que um termo candidato aparece como palavra-chave no conjunto de treinamento (KEYPHRASENESS);
- Relações semânticas: O grau do nodo (NODE DEGREE) quantifica o número de relações semânticas de um termo do tesauro correspondente a uma frase candidata com outros termos do tesauro correspondentes a outras frases candidatas do mesmo documento. Tenta capturar a proximidade semântica entre frases (do inglês *semantic relatedness*), já que geralmente, palavras-chaves tem alta coocorrência com termos relacionados incluídos num tesauro.

Cabe ressaltar que a qualidade dos recursos que são fornecidos para o algoritmo de aprendizado de máquina envolvido é a chave para o sucesso da indexação automática pelo MAUI, contudo, é necessário que se tenha conhecimento necessário para a utilização dos recursos no contexto da indexação, o que ressalta a importância do trabalho do profissional da informação.

4.3 ANÁLISE COMPARATIVA

Por meio da análise de conteúdo dos trabalhos selecionados como relatos de experiência no uso dos sistemas, foi possível analisar comparativamente as características descritivas do processo de indexação de cada software, levantando informações acerca de como são processados os documentos e como é feita a extração e seleção dos termos que resulta nos descritores propostos por cada sistema.

Para tanto, foram utilizadas categorias para descrever as características de cada sistema, cujos valores podem ser observados no Quadro 2.

Quadro 2 – Características dos sistemas SISA e MAUI

CARACTERÍSTICA	SISA	MAUI
Arquivos utilizados no processamento	-Texto completo: texto com marcações de título, resumo e texto; -Lista de stopwords; -Vocabulário controlado formatados em dois arquivos textuais	- Texto completo: texto sem marcações; - Lista de stopwords; - Vocabulário controlado; - Conjunto de treinamento envolvendo texto dos documentos e respectivas palavras-chave, utilizado para treinar o modelo de aprendizagem de máquina.
Formato dos arquivos de entrada	TXT	TXT, SKOS (vocabulário controlado).
Utilização de Linguagem Documentária	Obrigatório. O vocabulário controlado deve ser formatado em arquivos de texto de acordo com o padrão exigido pelo SISA.	Obrigatório para a indexação automática por atribuição. Incorpora vocabulários controlados no formato SKOS.
Flexibilidade	Extensível para outros idiomas e outras áreas do conhecimento por meio do vocabulário controlado e lista de stopwords.	Extensível para outros idiomas e outras áreas do conhecimento por meio do vocabulário controlado, lista de stopwords, software radicalizador e conjunto de treinamento. Pode otimizar seu desempenho para coleções específicas.
Etapas de processamento	O processo de análise do documento é efetuado através das etapas: - Preparação das fontes; - Análise do conteúdo: identificação de termos do vocabulário controlado que ocorrem no documento constituído por título, resumo e texto. - Valoração e ponderação dos termos: leva em conta critérios preestabelecidos de frequência e posição dos termos no documento para propor os termos de indexação.	Inicialmente, consiste no treinamento de um modelo de aprendizado de máquina supervisionado para propor termos de indexação. Posteriormente, o processo de análise do documento é efetuado através das etapas: - Geração de termos candidatos à indexação; - Cálculo de características para os termos candidatos. - Aplicação do modelo de indexação treinado na ponderação e seleção de termos de indexação.
Operações sobre o texto	- Análise léxica; - Remoção de Stopwords; - Vocabulário controlado	- Análise léxica; - Remoção de Stopwords; - Radicalização; - Extração de n-gramas; -Vocabulário controlado
Ponderação dos termos	Frequência, posição	Frequência, posição, tamanho, probabilidade de ser palavra-chave e relações semânticas.
Abordagem na indexação automática por atribuição	Não supervisionada por meio de sistema especialista. Abordagem estatística e linguística.	Supervisionada no treinamento de modelo de aprendizagem de máquina. Abordagem estatística e linguística.

Fonte: Dados da pesquisa.

A partir do Quadro 2 e da análise comparativa das principais características dos sistemas analisados, pode-se concluir que ambos, apesar de possuírem características em comum, apresentam especificidades que os diferenciam entre si.

Com relação à característica “Arquivos utilizados no processamento”, que corresponde às fontes informacionais exigidas por cada sistema, nota-se que os sistemas apresentam requisitos em comum como o texto completo a ser indexado, uma lista de palavras vazias, e vocabulário controlado. Porém o MAUI exige adicionalmente um conjunto de treinamento envolvendo o texto de alguns documentos e respectivas palavras-chave, utilizado para treinar o modelo de aprendizado de máquina que classificará os termos de outros documentos como sendo palavra-chave ou não.

Na característica “Formato de arquivos de entrada”, que corresponde ao formato de arquivos exigidos por cada sistema, os sistemas possuem uma diferenciação acentuada, pois o SISA comporta arquivos nos formatos TXT, enquanto o MAUI comporta além de TXT, o formato SKOS na especificação de vocabulário controlado.

Referente à característica “Utilização de linguagem documentária”, que corresponde à padronização da linguagem de indexação pelos sistemas, tanto o SISA quanto o MAUI apresentam a obrigatoriedade da entrada de um vocabulário controlado para realizar a indexação automática por atribuição. Ou seja, os sistemas atuam em conformidade com uma linguagem documentária. Entretanto, o MAUI também pode realizar a indexação automática por extração quando o vocabulário controlado não é apresentado, e o SISA exige o vocabulário controlado. O SISA apresenta a singularidade de exibir na tela para o indexador também termos candidatos que não constam no vocabulário.

Com relação à “Flexibilidade”, que corresponde à capacidade de os sistemas operarem em outros idiomas e domínios, ambos se apresentam extensíveis para outros idiomas e áreas de conhecimento por meio de vocabulário controlado e lista de palavras vazias, porém, o MAUI apresenta outros requisitos como: um software radicalizador de palavras para o idioma do texto; e um conjunto de treinamento, necessário para gerar um modelo de indexação para coleções específicas.

Referente à característica “Etapas de processamento”, que corresponde às etapas que cada sistema executa para indexar os textos, os sistemas apresentam diferenças. Onde o SISA necessita de três etapas para efetuar a análise do documento: Preparação das fontes; Análise do conteúdo; e Valoração e ponderação dos termos. Enquanto o

MAUI requer inicialmente, o treinamento de um modelo de aprendizado de máquina supervisionado para propor termos de indexação, para a posterior análise de documentos em três etapas: Geração de termos candidatos à indexação; Cálculo de características para os termos candidatos; e Aplicação do modelo de indexação treinado para selecionar os termos de indexação com maior probabilidade de serem palavras-chave.

Com relação às “Operações sobre o texto”, que são operações de processamento da linguagem natural que são aplicadas no pré-processamento dos textos com o objetivo de melhorar a qualidade do processamento das informações via extração de termos relevantes (BAEZA-YATES; RIBEIRO NETO, 2013), os sistemas apresentam operadores comuns como a análise léxica, eliminação de stopwords e uso de vocabulário controlado, sendo que o MAUI apresenta adicionalmente a extração de n-gramas e radicalização das palavras.

Acerca da “Ponderação dos termos”, que consiste na análise das características referente aos termos de indexação, os sistemas também apresentam similaridade como o uso de características baseadas na frequência e posição dos termos, se diferenciando o MAUI por incluir características como o tamanho do termo, a probabilidade de ser palavra-chave no conjunto de treinamento ou domínio e a quantidade de relações semânticas do termo candidato com outros termos candidatos no tesauro.

A última característica analisada foi a “Abordagem na indexação automática por atribuição”, que consiste na classificação do método de indexação automática levando em consideração a abordagem na incorporação do conhecimento do especialista e abordagem no processamento da linguagem natural, presentes em cada software. A pesquisa mostrou divergência quanto ao método de incorporação do conhecimento do especialista nos sistemas. No SISA, a abordagem é não supervisionada por meio de sistema especialista, onde regras heurísticas de indexação foram embutidas no software. Com relação ao MAUI, a abordagem é supervisionada pois envolve treinamento de modelo de aprendizagem de máquina. Os sistemas convergem quanto à abordagem estatística e linguística no processamento da linguagem natural. Em ambos, a abordagem linguística no processamento da linguagem natural ocorre devido ao uso do vocabulário controlado, adicionalmente o MAUI apresenta o uso de radicalização de palavras.

Por meio do quadro, fica evidente que SISA e MAUI possuem valores similares para algumas características, tais como quanto aos arquivos utilizados no processamento, formato dos arquivos de entrada, utilização de linguagem documentária, flexibilidade,

ponderação dos termos, e operações sobre o texto. Porém, o MAUI se destaca por possuir valores singulares adicionais em relação ao SISA nessas características.

Com relação aos demais critérios utilizados no quadro, percebe-se que o MAUI se diferencia em relação ao SISA quanto ao critério de abordagem na indexação automática por atribuição e por possuir características distintas quanto às etapas de processamento.

Cabe ressaltar que o MAUI apresenta uma característica específica de aprendizado de máquina, que pode ser um fator determinante para uma melhor eficácia na indexação automática para domínios específicos.

Para além do quadro comparativo das características dos sistemas, na literatura constata-se que o SISA foi aplicado na indexação automática de publicações científicas escritas em português em áreas do conhecimento como Odontologia, Agricultura e Ciência da Informação, com a utilização dos respectivos vocabulários controlados: DeCS, Thesagro e Tesauro Brasileiro de Ciência da Informação (TBCI). Os trabalhos apresentaram o objetivo de analisar o resultado da indexação automática por atribuição por meio do SISA, sendo eles: (NARUKAWA, GIL LEIVA, FUJITA, 2009); (NARUKAWA, 2011); (GIL LEIVA, 2017); (BANDIM; CORRÊA, 2018) e (BANDIM; CORRÊA, 2019).

Quanto ao MAUI, este não foi aplicado na indexação automática por atribuição de documentos científicos em português, apenas nos idiomas inglês, espanhol e francês. Tais experimentos se encontram descritos na tese de Medelyan (2009). O MAUI foi aplicado na indexação automática por atribuição de documentos científicos em inglês das áreas de agricultura, medicina e física, sendo utilizado os respectivos tesauros: Agrovoc, MeSH (*Medical Subject Headings*) e HEP (*High Energy Physics thesaurus*). Para documentos nos idiomas espanhol e francês foram realizados experimentos na área de agricultura, usando o Agrovoc como vocabulário controlado.

Embora os trabalhos discutidos acima reportem valores de métricas de qualidade na indexação, e em alguns relatos concluam que a qualidade da indexação dos sistemas se aproxima da realizada por indexadores, os valores não são diretamente comparáveis, nem permitem generalizações quanto à superioridade de um sistema em relação ao outro. Isto torna latente a necessidade de avaliação conjunta desses sistemas na indexação automática por atribuição em um mesmo conjunto de publicações científicas em língua portuguesa.

5 CONSIDERAÇÕES FINAIS

O desconhecimento acerca de características descritivas e comparativas de sistemas de indexação automática é um dos principais desafios na descrição, avaliação e utilização consciente dos mesmos nos ambientes informacionais. Por meio desta pesquisa, buscou-se discutir a indexação automática por atribuição, bem como apresentar os sistemas SISA e MAUI, levantar características descritivas quanto à operação e uso dos sistemas, além de descrever tais sistemas e depois comparar os valores das principais características.

Por meio da análise comparativa das características dos dois sistemas, foi possível perceber seus requisitos, recursos utilizados e a forma como cada sistema opera, mediante suas especificidades e exigências.

A análise comparativa resultante desta pesquisa permite aos indexadores e pesquisadores, o acesso às informações mais relevantes e precisas acerca dos sistemas analisados. Além disso, as características descritivas propostas neste trabalho podem ser reutilizadas na descrição de outros sistemas de indexação automática por atribuição, que poderão então ser comparados aos sistemas aqui descritos.

Especificamente quanto ao SISA e MAUI, os sistemas apresentam valores para as características em comum, porém o MAUI se mostra mais promissor por conta dos recursos adicionais de processamento de linguagem natural que implementa, e por apresentar a especificidade de treinamento de um modelo de indexação por meio de um algoritmo de aprendizado de máquina. O uso de aprendizado de máquina promete oferecer melhor eficácia na representação da informação por gerar modelo a partir da indexação manual, considerada a melhor forma de indexação por pesquisadores da área.

O artigo ao descrever o processamento dos sistemas SISA e MAUI, contribui também no acesso às informações necessárias para utilização dos sistemas, como requisitos de entrada e recursos utilizados na indexação automática.

Assim, espera-se que esta pesquisa contribua para a área da indexação automática, como meio de esclarecimento acerca das características dos sistemas e de sua utilização, bem como venha a fornecer subsídios para a elaboração de novas pesquisas.

Quanto ao contexto de aplicação dos sistemas, os mesmos já foram utilizados em diferentes domínios e alcançaram em alguns experimentos uma qualidade na indexação próxima à obtida por indexadores, porém é necessário analisar comparativamente a

qualidade da indexação realizada por ambos por meio de experimento para um mesmo conjunto publicações científicas escritas em português.

Aponta-se como possível desdobramento da presente pesquisa a avaliação da qualidade da indexação automática dos sistemas SISA e MAUI num mesmo corpus, por meio de uma análise comparativa dos índices de qualidade na indexação, a fim de validar a eficácia dos sistemas e apresentar possíveis aperfeiçoamentos quanto ao processo de indexação.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. Recuperação de informação: conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

BANDIM, M. A. S.; CORREA, R. F. Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação. **Transinformação**, v. 31, p. 1-12, 2019.

BANDIM, M. A. S; CORRÊA, R. F. A consistência na indexação automática por atribuição de artigos científicos na área de Ciência da Informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 64-77, set. 2018.

BORKO, H.; BERNIER, C. **Indexing concepts and methods**. New York: Academic Press, 1978.

CAJUEIRO, R. L. P. **Manual para elaboração de trabalhos acadêmicos**: guia prático do estudante. 3. ed. Petrópolis: Vozes, 2015. 110 p.

CERVO, A. L.; BERVIAN, P. A.; SILVA, R. **Metodologia científica**. 6. ed. São Paulo: Pearson Prentice Hall, 2007. 162 p.

CINTRA, A. M. M. Elementos de lingüística para estudos de indexação. **Ciência da Informação**, v. 12, n. 1, 1983.

FUJITA, M. S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**. Campinas, v. 1, n. 1, p. 60-90, jul/dez. 2003.

FUJITA, M. S. L. **A leitura documentária do indexador**: aspectos cognitivos e linguísticos influentes na formação do leitor profissional. 2003. 21 f. Tese (Livre-Docência em Análise Documentária e Linguagens Documentárias Alfabéticas) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2003.

FUJITA, M. S. L.; GIL-LEIVA, I. **As linguagens de indexação em bibliotecas nacionais, arquivos nacionais e sistemas de informação na América Latina**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2010.

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

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

GIL LEIVA, I. SISA – Automatic indexing system for scientific articles: Experiments with location heuristics rules versus TF-IDF Rules. **Knowledge Organization**, v.44, n. 3, p. 139-162, 2017.

HJØRLAND, B. Automatic Indexing. In: **Lifeboat for Knowledge Organization**, 2008.

KIM, S. N.; MEDELYAN, O.; KAN, M.Y.; BALDWIN, T. Automatic Keyphrase extraction from Scientific Articles. In**: Language Resources and Evaluation** (2013) v. 47, n. 3, p. 723-742, Springer. December 2012.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos Livros, 2004. 452p.

LIMA, V. N. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em ciência da informação do vocabulário controlado do sibi/usp nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, 2009.

MEDELYAN, O. **Human-competitive automatic topic indexing.** PhD Thesis. University of Waikato, New Zealand, 2009. Disponível em: https://hdl.handle.net/10289/3513 . Acesso em: 26/06/2019.

MEDELYAN, O.; FRANK, E.; WITTEN, I.H. Human-competitive tagging using automatic keyphrase extraction. In: **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing**, 2009.

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

NARUKAWA, C. M. **Estudo de Vocabulário Controlado na Indexação Automática:** Aplicação no Processo de Indexação do Sistema de Indización Semiautomatica (SISA). 2011. 222 f. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

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 e Sociedade: Estudos**, João Pessoa, v.19, n.2, p. 99-118, 2009.

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.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, v. 20, n. 2, 1991.

SILVA, M.R.; FUJITA, M.S.L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. Campinas**: Transinformação**. V. 16(2), p. 133-161, maio/ago. 2004.

NOTAS

AGRADECIMENTOS
Não se aplica.

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: S. R. B. Silva, R. F. Correa
Coleta de dados: S. R. B. Silva, R. F. Correa
Análise de dados: S. R. B. Silva, R. F. Correa
Discussão dos resultados: S. R. B. Silva, R. F. Correa
Revisão e aprovação: S. R. B. Silva, R. F. Correa

CONJUNTO DE DADOS DE PESQUISA

Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
Não se aplica.

CONFLITO DE INTERESSES
Não se aplica.

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO
Recebido em: 18/12/2019 – Aprovado em: 19/05/2020 – Publicado em: 10/07/2020