Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação

Automatic indexing by assignment of scientific articles written in Portuguese from the Information Science area

Marcio Aercio Silva BANDIM[1] 0000-0002-8205-244X

Renato Fernandes CORREA[1] 0000-0002-9880-8678

Resumo

Propõe e avalia um processo de indexação automática por atribuição na representação de artigos escritos em português, visando a construção de uma base de dados científica na área de Ciência da Informação no Brasil. Utiliza como metodologia, a pesquisa exploratória, bibliográfica e empírica. A parte empírica envolve a realização de um experimento na forma de estudo de caso. O experimento consistiu na aplicação do processo proposto em um *corpus* formado por 60 artigos científicos e avaliação da qualidade na indexação automática por meio dos índices de consistência, precisão, revocação e medida F, tendo como padrão de referência as palavras-chaves dos autores. No processo proposto foram utilizados o Tesauro Brasileiro em Ciência da Informação e o *software* SISA. Foram obtidos resultados satisfatórios quanto a qualidade na indexação automática: índice de consistência médio de 19%, precisão média de 30%, revocação média de 37% e medida F média de 30%. Os resultados da pesquisa mostram que o Tesauro tem forte influência nos resultados de uma indexação automática por atribuição, apesar das relações de termo geral terem pouco contribuído para a qualidade na indexação automática. Também, foram apontados fatores intervenientes na indexação automática.

Palavras-chave: Indexação automática. Indexação automática por atribuição. Tesauro. Periódico científico. Ciência da Informação.

Abstract

This work proposes and evaluates a process of automatic indexing by assignment in the representation of full-text articles written in Portuguese, in the context of construction of a scientific database in the area of Information Science in Brazil. It uses the exploratory, bibliographic and empirical research as a methodology. The empirical part takes base in the accomplishment of an experiment as a case study. The experiment consists of the application of the proposed process in a corpus composed of 60 scientific articles, as well as quality assessment in automatic indexing through indexes of consistency, precision, recall, and F-measure. The gold standard was the authors' keywords. The automatic indexing process uses the Brazilian Thesaurus of Information Science and SISA software. The satisfactory results were a consistency index average of 19%, an average precision of 30%, an average recall of 37%, and a mean F-measure of 30%. The analysis of the results shows the thesaurus has a strong influence on the results of an automatic indexing by

[1] Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Programa de Pós-Graduação em Ciência da Informação. Av. da Arquitetura, s/n., *Campus* Universitário, 50740-550, Recife, PE, Brasil. Correspondência para/*Correspondence to*: R.F. CORREA. *E-mail*: <renato.correa@ufpe.br>.

Artigo elaborado a partir da dissertação de M.A.S. BANDIM, intitulada "Indexação automática por atribuição de artigos científicos da área de Ciência da Informação". Universidade Federal de Pernambuco, 2017.

Recebido em 30 de janeiro de 2017, e aprovado em 10 de agosto de 2018.

Como citar este artigo/*How to cite this article*

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, e180004, 2019. http://dx.doi.org/10.1590/2318-0889201931e180004

assignment, although the general term's relations had poor contribution on the quality of the automatic indexing. In addition, we point out intervening factors in automatic indexing.

Keywords: *Automatic indexing. Automatic indexing by assignment. Thesaurus. Scientific journal. Information Science.*

Introdução

Fazer Ciência é produzir conhecimento científico. Possuir esse conhecimento é essencial para mudar os rumos da história da humanidade. Por exemplo, graças aos conhecimentos produzidos pela humanidade, grandes epidemias hoje são evitadas, epidemias que por mais de uma vez devastaram populações inteiras em épocas passadas (Trzesniak, 2014).

Todavia, para o efetivo uso, o conhecimento precisa ser divulgado por meio da comunicação científica. Entre os instrumentos que potencializam a comunicação científica, se destacam as bases de dados de publicações científicas. Essas bases realizam o processo de indexação dos assuntos das publicações científicas, objetivando o acesso e recuperação das informações armazenadas.

Segundo Guimarães (2003), para disponibilizar o acesso à informação é preciso levar em consideração todas as questões que tratam dos processos de produção, coleta, organização e recuperação dessa informação.

A organização e recuperação da informação se materializa via indexação. A indexação é um dos processos de análise documentária realizada com a finalidade de determinar para cada documento um conjunto de palavras-chave ou assuntos. A indexação facilita a armazenagem em bases de dados e posterior recuperação dos documentos, atendendo desse modo as necessidades de recuperação da informação dos usuários (Fujita; Gil Leiva, 2014).

No contexto das bases de dados científicas, a indexação automática tem sido adotada visando dar conta da indexação do volume crescente de artigos científicos e da necessidade de criação de índices para busca. 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.

Segundo Gil Leiva (1999), a indexação automática faz uso de programas que realizam a análise do texto e propõem termos de indexação, sendo os termos propostos armazenados como descritores do documento, não necessitando de validação dos termos selecionados.

Existem basicamente dois tipos de indexação automática segundo Lancaster (2004): a indexação automática por extração – que extrai e seleciona de forma automática termos do texto dos documentos; e a indexação automática por atribuição – em que termos são atribuídos automaticamente de um vocabulário controlado, dado previamente para cada termo um conjunto de palavras ou expressões que ocorrem com frequência nos documentos.

No processo de indexação automática por atribuição, o vocabulário controlado atua no processo de análise automática do texto do documento e na sua representação, ou seja, condiciona os resultados da análise do conteúdo temático dos documentos e da atribuição de descritores (Narukawa; Gil Leiva; Fujita, 2009).

Considerando o cenário de construção de uma base de dados de texto completo de artigos de periódicos em Ciência da Informação no Brasil, tendo o português como idioma principal dos documentos nessa base, o presente artigo apresenta como problema de pesquisa a seguinte questão: como realizar a indexação automática por atribuição de artigos de periódicos escritos em português da área em Ciência da Informação?

Justifica-se a escolha de pesquisar a indexação automática por atribuição, por essa permitir o controle terminológico e o uso de tesauro, instrumento já utilizado na construção de sistemas de recuperação da informação especializados, bem como na interface de busca e acesso às bases de dados cientificas.

Assim, o objetivo deste trabalho é propor e avaliar um processo de indexação automática por atribuição na representação de artigos científicos em português da área de Ciência da Informação.

Trabalhos Relacionados

Analisando a literatura científica nacional e internacional na área de Ciência da Informação, foram considerados como trabalhos relacionados aqueles que aplicaram um processo de indexação automática por atribuição a documentos científicos escritos em português. Consequentemente, nessa subseção são discutidos os seguintes trabalhos relacionados: (Lima; Boccato, 2009); (Narukawa; Gil Leiva; Fujita, 2009); (Souza; Gil Leiva, 2016) e (Gil Leiva, 2017).

Lima e Boccato (2009) utilizaram o *software* SISA na indexação automática por atribuição de resumos de teses e dissertações escritos em português da área de Ciência da Informação. O objetivo do artigo foi avaliar o desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do Sistema Integrado de Bibliotecas/Universidade de São Paulo (SIBI/USP) nos processos de indexação manual, automática e semiautomática de um *corpus*. O *corpus* consistiu de 70 resumos de dissertações e teses em português cadastradas no Banco de Dados Bibliográficos da USP (Dedalus), abrangendo o período de janeiro de 2002 a dezembro de 2007.

Diferentemente do presente trabalho, Lima e Boccato (2009) não utilizaram artigos científicos nem o texto completo dos documentos, não utilizaram um vocabulário controlado especializado na área de Ciência da Informação, e não reportaram os valores alcançados para métricas de qualidade na indexação automática.

Outro trabalho relacionado é o relato de pesquisa de Narukawa, Gil Leiva e Fujita (2009), que analisou a aplicação do *software* SISA com uso da terminologia do Descritores em Ciência da Saúde (DeCS) na indexação automática por atribuição de artigos de periódico escritos em português da área de Odontologia. Foi realizada uma análise comparativa entre a indexação automática do SISA e a indexação manual do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (Bireme). Foram reportadas as métricas de consistência na indexação, bem como revocação e precisão na recuperação da informação.

O presente trabalho adota as métricas de qualidade de indexação automática reportadas na pesquisa de Narukawa, Gil Leiva e Fujita (2009), se diferenciando no domínio do tesauro e dos artigos de periódicos alvos de indexação, e na forma de cálculo das medidas de revocação e precisão.

Em (Souza; Gil Leiva, 2016) e (Gil Leiva, 2017) são reportados experimentos com aplicação do *software* SISA e vocabulário controlado Thesagro na indexação automática de artigos em português da área de Fruticultura. São reportadas métricas de qualidade na recuperação da informação para um conjunto de expressões de busca a fim de mensurar a qualidade na indexação automática. Esses trabalhos diferem do presente artigo por utilizarem artigos e vocabulário controlado da área de fruticultura. As métricas reportadas de qualidade na recuperação da informação também diferem das métricas de qualidade na indexação automática utilizadas no presente trabalho.

Todavia, percebe-se que todos os trabalhos relacionados utilizaram o *software* SISA como base dos diferentes processos de indexação automática. No entanto nenhum deles propôs e avaliou um processo de indexação automática por atribuição para artigos científicos da área de Ciência da Informação, como apresentado no presente artigo.

Métodos

Quanto aos objetivos, a metodologia desta pesquisa é de natureza exploratória, e quanto aos métodos faz uso da pesquisa bibliográfica e da pesquisa empírica.

Foi realizada primeiramente uma pesquisa bibliográfica, consistindo da análise de trabalhos relacionados da literatura nacional e internacional da área de Ciência da Informação sobre indexação automática por atribuição de documentos escritos em português. Por meio da pesquisa bibliográfica buscou-se determinar: um *software* que realizasse a indexação automática por atribuição para textos escritos em português; um tesauro da área de Ciência da Informação com termos em português; uma metodologia para avaliação da qualidade na indexação automática; e elementos para a proposição de um processo de indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação, levando em conta os processos aplicados nos trabalhos relacionados.

A parte empírica da pesquisa envolveu um experimento onde foi aplicado o processo proposto de indexação automática por atribuição a um *corpus* de artigos científicos.

O experimento foi realizado nos moldes de um estudo de caso, consistindo da indexação automática por atribuição de documentos de um *corpus* composto por 60 artigos científicos referenciados na tese de doutorado de Souza (2005). Os artigos do *corpus* consistem de documentos predominantemente textuais, escritos em língua portuguesa, publicados nos anos de 2002 e 2003 em duas revistas científicas, eletrônicas da área da Ciência da Informação: DataGramaZero e Ciência da Informação.

Na composição da proposta de processo de indexação automática por atribuição, foi utilizado o *software* SISA. Justifica-se essa escolha por esse *software* ser utilizado na unanimidade dos trabalhos relacionados (vide subseção trabalhos relacionados).

Corroboram com esse resultado o trabalho de Lapa e Correa (2014), que apontam os seguintes *softwares* mais aplicados nos estudos brasileiros em Ciência da Informação sobre a temática indexação automática:

1) O sistema PRECIS – utilizado na construção automática de índices de assunto em (Fujita, 1989);
2) O AUTOMINDEX/II – aplicado na geração de tabela de frequências de aparecimento de descritores em (Robredo, 1991);
3) O SISA – utilizado na indexação automática por atribuição nos trabalhos reportados na subseção de trabalhos relacionados.
4) O OGMA – aplicado na indexação automática por extração de sintagmas nominais em (Maia; Souza, 2010; Corrêa *et al*., 2011; Corrêa; Bazilio, 2017).

Dentre os sistemas apontados, o *software* SISA é o único que permite a realização da indexação automática por atribuição, tendo sido cedido pelo seu desenvolvedor para fins de desenvolvimento deste artigo.

Adicionalmente, para composição da proposta de processo de indexação automática por atribuição, torna-se necessário a adoção de um tesauro de especialidade no idioma dos documentos. Após pesquisa bibliográfica e análise dos principais tesauros da área de Ciência da Informação nacionais e internacionais, optou-se por utilizar como vocabulário controlado o tesauro definido em (Pinheiro; Ferrez, 2014), denominado Tesauro Brasileiro em Ciência da Informação (TBCI), por ser o mais atual e contemplar termos em português.

Visando a avaliação da qualidade na indexação automática obtida por meio do processo proposto de indexação automática, foram comparadas as palavras-chave dos artigos com os termos do TBCI atribuídos pelo software SISA na indexação automática por atribuição. Os termos de indexação propostos pelo sistema foram dispostos numa planilha eletrônica onde constam também as palavras-chave dos respectivos artigos. Assim, foi possível comparar os dois conjuntos de termos de indexação, marcar os termos comuns que casam totalmente (negrito) e parcialmente (negrito e itálico). O Quadro 1 ilustra uma amostra da planilha eletrônica criada para comparação dos termos de indexação.

No experimento, são realizadas as avaliações intrínseca e extrínseca do processo de indexação automática. A avaliação intrínseca mede o grau de consistência da indexação e a avaliação extrínseca mede o grau de Revocação, Precisão e Medida F na indexação automática.

Quadro 1. Amostra da planilha para comparação dos termos de indexação.

Artigo Científico	Termos das Palavras-chave	Termos de Indexação do SISA
Artigo 1	1 - Transferência de Informação	1 - Transferencia da Informação
	2 - Gestão do Conhecimento	2 - Avaliação
	3 - Valor de Unidades de Conhecimento	3 - Acesso
		4 - Descarte
		5 – Estudos de Caso
		6 - Gestão
		7 - Gestão do Conhecimento
Artigo 2	1 - Popularização da Ciência	1 - Comunicação Cientifica
	2 - Comunicação Científica	2 - Ciencia da Informação
		3 - Estudos de Caso
		4 - Educação
		5 - Notícias
Artigo 3	1 - Informação	1 - Avaliação
	2 - Valor Informacional	2 - Direito
	3 - Direito à Informação	3 - Direito à Informação
	4 - Memória Social	4 - Recuperação da Informação
	5 - Estoque Informacional	

Fonte: Elaborado pelos autores (2017).

A avaliação intrínseca quantitativa mede o grau de consistência da indexação automática com a indexação intelectual dos autores (palavras-chave dos artigos).

De acordo com Gil Leiva (1997, p.22), a consistência da indexação é definida, como o grau de concordância entre indexadores de um mesmo grupo ou entre indexadores de grupos diferentes, quando da representação da informação essencial de um documento, por meio de um conjunto de termos de indexação selecionados por esses indexadores.

Na análise de consistência, são comparados os termos de indexação das palavras-chave dos autores dos artigos científicos (indexação A), com os termos de indexação atribuídos pelo SISA (indexação B).

A Equação (1) foi usada para o cálculo do índice de Consistência (C). Onde: Tco=Número de termos comuns nas duas indexações; A=Número de termos usados na indexação A; e B=Número de termos usados na indexação B.

$$C = \frac{Tco}{(A + B) - Tco} \qquad \text{Equação (1)}$$

Adicionalmente, uma indexação é considerada de boa qualidade quando os termos de indexação disponibilizados aos usuários representam os itens informacionais atribuindo a totalidade de termos relevantes sem incluir termos pouco relevantes como descritores documentais. Assim, é necessário fazer uma avaliação extrínseca na atribuição automática aos documentos das palavras-chaves dos autores (consideradas como termos relevantes), por meio do cálculo das métricas de Revocação, Precisão e Medida F.

O índice de Revocação (R) é obtido por meio da relação entre os termos relevantes atribuídos e o total de termos relevantes existente para cada artigo, como mostra a Equação (2).

$$R = \frac{\text{Número de termos relevantes atribuídos}}{\text{Número total de termos relevantes}}$$ Equação (2)

Para se obter o índice de Precisão (P), calcula-se a relação entre os termos relevantes atribuídos e o total de termos atribuídos para cada artigo, vide a Equação (3).

$$P = \frac{\text{Número de termos relevantes atribuídos}}{\text{Número total de termos atribuídos}}$$ Equação (3)

A medida F é a média harmônica entre o índice de Precisão e o índice de Revocação, sendo uma maneira de combinar a Precisão e Revocação em um único número. Se nenhum termo relevante for recuperado, a Medida F assume valor zero e, quando todos os termos atribuídos forem relevantes (máxima Precisão) e forem exaustivamente recuperados todos os termos relevantes (máxima Revocação), assume valor igual a um que corresponde a 100%. A medida F só assume valores altos quando os índices de Revocação e Precisão são altos também, refletindo um *equilíbrio* entre os dois índices antagônicos. A equação (4) é usada para o cálculo da Medida F.

$$F = \frac{(2 \times (P \times R)}{(P + R)}$$ Equação (4)

O critério de consistência relaxada (Narukawa, 2011) foi utilizado para medir a equivalência entre termos atribuídos e palavras-chave, e posteriormente foram contabilizados os termos comuns (ou termos relevantes atribuídos) e calculadas as métricas de avaliação da qualidade na indexação automática.

O Quadro 2 ilustra uma amostra da planilha eletrônica criada para contabilização dos termos de indexação e cálculo das métricas de qualidade na indexação automática.

A avaliação da indexação automática foi realizada em dois cenários: Cenário I – com a lista de termos gerais habilitada na configuração do SISA; e Cenário II – sem habilitar essa lista. Isto se justifica devido a habilitação da atribuição de termos gerais ter influência nos índices de consistência, Revocação, Precisão e Medida F.

A próxima subseção apresenta a proposta de processo de indexação automática por atribuição.

Proposta de processo de indexação automática

A presente proposta de processo de indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação tem por base a aplicação do *software* SISA com o uso do tesauro TBCI.

Para o uso do *software* SISA são necessários como entrada arquivos contendo respectivamente: arquivos extraídos de um vocabulário controlado no mesmo idioma do documento; uma lista de palavras vazias no idioma do documento; e um documento a indexar em formato de texto com marcadores de campo (título, resumo e texto completo).

Quadro 2. Amostra da planilha para cálculo das métricas de qualidade.

Artigo Científico	Palavras-chave	Termos do SISA	Termos comuns	Consistência	Precisão	Revocação	Medida F
	n			(%)			
Artigo 1	3	7	2	25	29	67	40
Artigo 2	2	5	1,5	17	20	50	29
Artigo 3	5	4	1,5	20	38	30	33

Fonte: Elaborado pelos autores (2017).

http://dx.doi.org/10.1590/2318-0889201931e180004

Assim, o processo proposto de indexação automática consiste de duas etapas: preparação dos arquivos de entrada do SISA; e a indexação automática dos documentos.

Na primeira etapa foram elaborados os seguintes arquivos de entrada para o *software* SISA:

a) Linguagem de indexação: constando de dois arquivos elaborados a partir do TBCI em formato de arquivo de texto, sendo: Arquivo I – lista de descritores – arquivo contendo os termos do TBCI em ordem alfabética, enumerados sequencialmente, onde os descritores estabelecem apenas a relação de equivalência e termo preferencial por meio da indicação USE; Arquivo II – lista de termos gerais – arquivo contendo a lista de termos com respectivo termo geral por meio da indicação TG, consistindo da extração das relações do tipo TG entre termos do TBCI.

b) Lista de palavras vazias: arquivo contendo uma lista de palavras vazias em língua portuguesa. Foi adaptada uma lista de *stopwords* cedida pelo coordenador do projeto de pesquisa intitulado "Mapeador Temático de Teses e Dissertações", na qual foram removidas as palavras isoladas que faziam parte da lista de descritores do TBCI.

c) Artigos científicos: arquivos contendo respectivamente o texto completo de cada artigo científico formatado para entrada no *software* SISA. Os arquivos foram obtidos da seguinte forma: os arquivos originais dos artigos foram convertidos para o formato de arquivo de texto utilizando o *software* Adobe Reader e Microsoft Word; as referências dos artigos foram removidas; o título, o resumo e o texto completo foram delimitados respectivamente por marcadores – #CTI# (começo do título) e #FTI# (fim do título); #CR# (começo do resumo) e #FR# (fim do resumo); #CTE# (começo do texto) e #FTE# (fim do texto).

Para evitar problemas na comparação dos termos pelo SISA, decorrentes da presença de caracteres especiais, todos os caracteres nos arquivos de entrada foram convertidos para caixa alta.

Na segunda etapa do processo, os arquivos de entrada foram abertos no *software* SISA e foi solicitada a realização da indexação automática por atribuição de cada um dos documentos.

Resultados e Discussão

Nesta seção são discutidos os resultados obtidos no experimento de avaliação da qualidade na indexação automática por atribuição dos artigos do *corpus*.

No Cenário I, com a lista de termos gerais habilitada na configuração do SISA, obteve-se para o *corpus* os resultados sintetizados no Quadro 3.

Na análise dos termos presentes no campo palavras-chave, houve um mínimo de 2 e um máximo de 9 termos nas palavras-chaves, dando uma média de 4 a 5 termos atribuídos por documento pelos autores dos trabalhos.

Quadro 3. Síntese dos resultados para o Cenário I.

	Palavras-chave	Termos do SISA	Termos comuns	Consistência	Precisão	Revocação	Medida F
	n			(%)			
Mínimo	2	3	0	0	0	0	0
Média	4,5	9,9	1,8	15	20	42	25
Desvio	1,6	4,9	1,0	8	12	23	12
Máximo	9	24	5	42	50	100	59

Fonte: Elaborado pelos autores (2017).

Com relação aos termos propostos pelo SISA, houve um mínimo de 3 e um máximo de 24 descritores atribuídos, e uma média de 9 a 10 termos de indexação atribuídos automaticamente para cada documento.

O número de palavras-chaves e número de termos atribuídos pelo SISA influenciam diretamente no valor dos índices de consistência e precisão por documento, uma vez que a quantidade de descritores atribuídos em cada tipo de indexação é considerada nos cálculos dessas métricas. Assim, as médias do número de termos nas palavras-chaves e do número de termos atribuídos pelo SISA estão correlacionadas com as médias de consistência e precisão para todo o *corpus*.

O número de termos comuns variou de um mínimo de zero termos comuns a um máximo de 5 termos comuns. Sendo a média de um ou dois termos comuns entre as indexações por documento.

Foi obtida uma média de 15% no índice de consistência, com uma variação de um mínimo de 0% a um máximo de 42%. Quanto aos índices de Revocação, Precisão e Medida F, seguem os seguintes resultados: Precisão média de 20%; Revocação média de 42% e Medida F média de 25%.

No Cenário II, ao desabilitar o arquivo de termos gerais e executar novamente a indexação automática dos 60 artigos, percebeu-se que os termos propostos pelo SISA diminuíram em quantidade, afetando diretamente os índices de consistência na indexação, Precisão, Revocação e Medida F. Os resultados para o *corpus* no Cenário II são sintetizados no Quadro 4.

Como era esperado, o número de palavras-chave se manteve inalterado para o *corpus* no Cenário II.

Com relação aos termos propostos pelo SISA, houve um mínimo de 1 e um máximo de 11 descritores atribuídos, e uma média de 5 termos de indexação atribuídos automaticamente por documento.

De um cenário para outro, o número médio de termos propostos pelo SISA foi reduzido pela metade, enquanto o número médio de termos comuns às duas indexações permaneceu próximo (1,8 no Cenário I e 1,5 no Cenário II, com desvio padrão de 1,0 em ambos).

Os índices de consistência foram "otimizados" com o SISA desabilitado a atribuir os termos gerais. Nesse segundo cenário, a média da consistência foi de 19%, com um mínimo de 0% e um máximo de 71% e, as médias dos outros índices foram: Revocação média de 37%; Precisão média de 30% e Medida F média de 30%.

Comparando-se os Quadros 3 e 4, pode-se observar como a diminuição da quantidade de termos atribuídos pelo SISA, decorrente da não habilitação da atribuição dos termos gerais, influi diretamente nos valores médios dos índices. A habilitação da atribuição de termos gerais no SISA causa uma diminuição dos índices médios de consistência (em 4%), precisão (em 10%) e medida F (em 5%), em detrimento de um aumento de 5% na média de revocação. Sugere-se, em virtude disto, desabilitar a atribuição de termos gerais no SISA, pois melhora os valores médios para o índice de consistência e medida F, sendo uma configuração mais equilibrada ou com melhor compromisso na atribuição de termos do TBCI equivalentes às palavras-chaves.

Corrobora também com essa sugestão o fato de que a coincidência entre os termos gerais atribuídos pelo SISA com os termos das palavras-chave foi mínima: do total de 268 palavras-chave do *corpus* de 60 artigos científicos, menos de 14 termos gerais atribuídos pelo SISA (5%) seriam equivalentes às palavras-chave.

Quadro 4. Síntese dos resultados para o Cenário II.

	Palavras-chave	Termos do SISA	Termos comuns	Consistência	Precisão	Revocação	Medida F
	n			(%)			
Mínimo	2	1	0	0	0	0	0
Média	4,5	5,1	1,5	19	30	37	30
Desvio	1,6	2,1	1,0	13	20	25	18
Máximo	9	11	5	71	100	100	83

Fonte: Elaborado pelos autores (2017).

 http://dx.doi.org/10.1590/2318-0889201931e180004

Segundo Narukawa (2011), a atribuição do termo geral é interessante pois confere à indexação maior revocação na recuperação da informação. No entanto, no contexto da indexação automática, a regra do SISA de inclusão de termos gerais para cada termo específico atribuído não contribui para a qualidade da indexação automática.

Em termos de comportamento médio do SISA, os valores médios das métricas para o melhor cenário (Cenário II) podem ser interpretados da seguinte forma: a consistência média de 19% significa que aproximadamente 2 termos de indexação são comuns entre cada 10 termos resultantes da união dos termos atribuídos pelo *software* e as palavras-chave de cada documento; a Revocação média de 37% significa que pouco mais de um terço das palavras-chave foram atribuídas pelo software para cada documento; a precisão média de 30% significa que quase um terço dos termos atribuídos pelo *software* são palavras-chave dos documentos.

Os valores médios alcançados para o melhor cenário apontam um desempenho satisfatório do processo de indexação automática por atribuição proposto neste trabalho.

Apesar das diferenças no domínio dos documentos e do tesauro, e no cálculo das métricas de Revocação e Precisão, os valores obtidos no presente trabalho para as métricas de qualidade na indexação automática por atribuição são próximos aos obtidos por Narukawa, Gil Leiva e Fujita (2009) para o domínio de Odontologia (a saber, índice médio de Consistência de 23,25%, índice médio de Precisão de 40,92%, e índice médio de Revocação de 35,72%).

Assim, como em (Narukawa, 2011), foram encontrados fatores intervenientes que impossibilitaram a atribuição pelo SISA das palavras-chave aos documentos.

No Quadro 5, são exemplificados fatores intervenientes observados durante o experimento. Esses fatores decorrem das regras internas de atribuição de termos pelo SISA, que atribui os termos que constam no vocabulário controlado e que se encontram no texto do documento, selecionando os melhores candidatos por meio de regras de frequência e recorrência desses termos no título, resumo e texto completo do artigo analisado.

Quadro 5. Fatores intervenientes no processo de indexação por atribuição.

Fatores Intervenientes	Termos das Palavras-chave	Termos do SISA
Termos no singular e no plural	- Biblioteca virtual	- Bibliotecas virtuais
	- Estratégia de busca	- Estratégias de busca
	- Biblioteca híbrida	- Bibliotecas híbridas
Dificuldade em atribuir termos compostos	- Compressão semântica	- Semântica
	- Ciência da informação no Brasil	- Ciência da informação
	- *Internet* e produção científica	- *Internet*
Diferenças na estrutura dos termos de indexação	- Recuperação de informação	- Recuperação da informação
	- Profissional da informação	- Profissionais de informação
	- Transferência de informação	- Transferência da informação
Dificuldade em atribuir conceitos implícitos	- Arquivos-abertos	- Acesso
	- Sistema de publicação	- Tipos de documentos
	- Interação humano-computador (IHC)	- Buscas de informação
Atribuição de termo específico a termo geral	- Informação	- Direito à informação
	- Ciência	- Ciência da Informação
	- Caos	- Teoria do caos

Fonte: Elaborado pelos autores (2017).

Foram encontrados os seguintes fatores intervenientes: termos no singular e plural; dificuldade em atribuir termos compostos; diferença na estrutura dos termos, ou seja, termos escritos de forma diferente (com omissão ou não de artigos, entre aspas, hifens e parênteses); dificuldade em atribuir conceitos implícitos e, atribuição de termo específico a termo geral.

No caso de termos no singular e no plural foram constatadas variações de um mesmo termo, como por exemplo, "Biblioteca virtual" e "Bibliotecas virtuais". Assim, o termo da palavra-chave do artigo não foi atribuído pelo SISA, mas uma variação no plural em alguns poucos documentos.

Outro fator interveniente foi a dificuldade do SISA em atribuir termos compostos. Por exemplo, em um documento constava como palavra-chave o termo composto "Compressão semântica", mas no texto aparecia com frequência a palavra "semântica". Assim, o SISA atribuiu o termo "semântica" ao documento.

Foram percebidas diferenças na estrutura dos termos de indexação como um fator interveniente. Essa diferença estrutural ocorreu principalmente no uso de preposição ou contração de preposição e artigo para unir substantivos nos termos. Por exemplo, em um documento a palavra-chave era "Recuperação de informação", e o termo "Recuperação da informação" foi atribuído pelo SISA.

Adicionalmente, o SISA teve dificuldade em atribuir conceitos implícitos relacionados. Por exemplo: nas palavras-chave constava "Arquivos-abertos" e, o SISA, atribuiu o termo "Acesso".

Por fim, foi notado como fator interveniente a atribuição de termo específico a termo geral. Nesse caso, o termo geral atribuído pelo autor foi o termo "Caos", e o SISA atribuiu o termo específico "Teoria do caos".

Portanto, os fatores intervenientes do Quadro 5 estão associados às questões que caracterizam as diferenças morfológicas, sintáticas e semânticas dos termos da indexação intelectual e da indexação automática utilizando vocabulário controlado. Esses fatores ocorrem porque o vocabulário controlado condiciona a atribuição dos termos de indexação ao documento pelo *software* SISA. Dessa forma, pode-se afirmar que o TBCI teve forte influência na qualidade da indexação alcançada pelo SISA para os documentos do *corpus*.

Outro tipo de fator interveniente observado foi a atribuição frequente pelo SISA de termos comuns, como o termo "Pesquisa" que foi atribuído a 18 documentos e "Ciência da Informação" que foi atribuído a 22 documentos. Todavia, esses termos se constituem palavras-chave somente de dois e dez documentos respectivamente. Isto influi diretamente na qualidade da indexação automática do SISA e, sugere, a necessidade de tratamento especial para termos do TBCI que são muito frequentes no texto dos artigos científicos.

Conclusão

Analisando-se os resultados deste trabalho, conclui-se que a proposta de processos de indexação automática por atribuição, com base no uso do *software* SISA com o tesauro TBCI, atende de forma satisfatória quanto a qualidade na indexação automática de artigos científicos em português da área de Ciência da Informação.

Corrobora com essa conclusão, os percentuais médios dos índices de Consistência, Precisão e Revocação encontrados para os documentos do *corpus*, que se encontram no mesmo patamar dos valores reportados na literatura para a indexação automática por atribuição de artigos científicos na área de Odontologia.

Também, é possível concluir que a habilitação da atribuição de termos gerais tem forte influência na indexação feita pelo *software* SISA, degradando os valores médios obtidos para os indicadores de qualidade na indexação automática como índice de Consistência, Precisão, e Medida F.

Entretanto, apesar da pouca contribuição das relações de termos gerais na qualidade da indexação automática alcançada pelo SISA, é perceptível a forte influência do vocabulário do tesauro no resultado da indexação automática por atribuição.

Durante a avaliação da qualidade na indexação automática por atribuição, foram encontrados fatores intervenientes que apontam para possíveis aprimoramentos nos algoritmos do SISA, visando uma melhor qualidade na indexação automática por atribuição com o uso de vocabulário controlado.

Como uma limitação deste trabalho, verificou-se que entre os 60 artigos analisados existem alguns com apenas duas palavras-chave e outros com mais de oito palavras-chave. Essa limitação pode ser resolvida reindexando intelectualmente os trabalhos com o uso do TBCI com um número fixo de palavras-chave para cada artigo.

Em termos de trabalhos futuros apontam-se: a reindexação dos trabalhos do *corpus* com o uso do TBCI; analisar a aplicação de ontologia como linguagem de indexação no lugar de tesauro; propor inclusão nos algoritmos do SISA de regras que levem em consideração os aspectos morfológicos, sintáticos e semânticos dos termos de indexação; e pesquisar a adaptação do processo proposto a outros *softwares* multilíngues de indexação automática por atribuição.

Mesmo com um desempenho satisfatório do processo de indexação automática por atribuição proposto, entende-se que na construção de bases de dados científicas deva haver o uso conjunto dos termos advindos da indexação manual e da indexação automática. Isto porque se constituem em conjuntos não disjuntos de termos atribuídos com perspectivas complementares, refletindo respectivamente a intenção dos autores e o texto dos documentos.

Com base nos dados obtidos com esta pesquisa, verificam-se pertinentes contribuições como a proposta de um processo de indexação automática por atribuição para artigos científicos em Ciência da Informação escritos em português do Brasil. Além da avaliação experimental desse processo proposto, possibilitando uma análise qualitativa do desempenho e a comparação posterior com variações desse processo, envolvendo outros *softwares* e vocabulários controlados.

Colaboradores

Todos os autores contribuíram na concepção e desenho do estudo, análise de dados e redação final.

Referências

Corrêa, R.F. *et al*. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. *AtoZ*, v.1, n.1, p.11-22, 2011.

Corrêa, R.F.; Bazilio, L.H.T. Análise da extração de descritores como sintagmas nominais através do software Ogma. *Encontros Bibli*: *Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v.22, n.50, 2017.

Fujita, M.S.L. Avaliação da eficácia de recuperação do sistema de indexação Precis. *Ciência da Informação*, v.18, n.2, p.120-134, 1989.

Fujita, M.S.L.; Gil Leiva, I. Avaliação da indexação por meio da recuperação da informação. *Ciência da Informação*, v.43, n.1, p.50-66, 2014.

Gil Leiva, I. *La automatización de la indización, propuesta teórico-metodológica*: aplicación al área de Biblioteconomía y Documentación. 1997. 268f. Tese (Doctorado) – Universidad de Murcia, Murcia, España, 1997.

Gil Leiva, I. *La automatización de la indización de documentos*. Gijón: Trea, 1999.

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.

Guimarães, J.A.C. A análise documentária no âmbito do tratamento da informação: elementos históricos e conceituais. In: Rodrigues, J.M.; Lopes, I.L. (Org.). *Organização e representação do conhecimento na perspectiva da Ciência da informação*. Brasília: Thesaurus, 2003. (Estudos Avançados em Ciência da Informação; 2). p.100-117.

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

Lapa, R.; Corrêa, R.F. Indexação automática no âmbito da Ciência da Informação no Brasil. *Informação & Tecnologia*, v.2, n.1, p.1-18, 2014.

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.

Maia, L.C.G.; Souza, R.R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência da Informação*, v.15, n.1, p.154-172, 2010.

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*, v.19, n.2, p.99-118, 2009.

Narukawa, C.M. *Estudo de vocabulário controlado na indexação automática*: aplicação no processo de indexação do Sistema de Indización Semiautomatica (SISA). 2011. 222f. Dissertação (Mestrado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

Pinheiro, L.V.R.; Ferrez, H.D. *Tesauro Brasileiro de Ciência da Informação*. Rio de Janeiro: IBICT, 2014.

Robredo, J. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. *Ciência da Informação*, v.11, n.1, p.3-18, 1991.

Robredo, J. *Documentação de hoje e de amanhã*: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivistas e museológicas. 4. ed. Brasília: Edição de autor, 2005.

Souza, R.R. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais*. 2005. 215f. Tese (Doutorado) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

Souza, R.R.; Gil Leiva, I. Automatic indexing of scientific texts: A methodological Comparison. *In*: *International Society for Knowledge Organization Conference*, 40., Rio de Janeiro, Brazil. Proceedings [...] Würzburg: Ergon Verlag, 2016. p.243-250.

Trzesniak, P. Indicadores quantitativos: como obter, avaliar, criticar e aperfeiçoar. Navus: *Revista de Gestão e Tecnologia,* v.4, n.2, p.5-18, 2014.