

PRINCÍPIOS PARA UTILIZAÇÃO DE ONTOLOGIAS NA INDEXAÇÃO AUTOMÁTICA

Maria Elisa Valentim Pickler – UNESP/MARÍLIA
Edberto Ferneda – UNESP/MARÍLIA

Resumo

O uso de ontologia como vocabulário controlado no processo de indexação permite agregar a esse processo uma terminologia de um domínio específico e uma estrutura conceitual que pode ser utilizada para realizar inferências, cujas relações permitem uma expansão dos termos extraídos por métodos puramente matemáticos. Defendemos a utilização das ontologias como vocabulário controlado para indexação automática. Ainda com o sentido que um tesouro tem para a indexação manual, de padronizar termos e estabelecer relações entre eles, mas além disso. As ontologias são legíveis por computador e permitem muitos tipos de relacionamentos. As ontologias abrem novas perspectivas para as pesquisas em indexação automática, uma vez que oferecem uma estrutura conceitual e terminológica restrita a um determinado domínio e originalmente representadas em linguagens legíveis por computador, o que permite a sua utilização nos mais variados processos computacionais.

Palavras-chave: Indexação automática. Ontologias. Vocabulários controlados.

1 INTRODUÇÃO

A indexação de um documento visa representar o seu conteúdo temático por meio de um conjunto de termos com o objetivo de sintetizar o seu conteúdo, ressaltando o que lhe é essencial. Os termos de indexação servem também como pontos de acesso mediante os quais o documento é localizado e recuperado em um sistema de informação.

As dificuldades inerentes à indexação manual e a grande quantidade de documentos publicados e disponibilizados, que caracterizou a “explosão informacional”, justificaram estudos que buscavam soluções alternativas para auxiliar o indexador no exercício de sua atividade. As primeiras pesquisas em indexação automática datam do final dos anos de 1950, época de rápido desenvolvimento das tecnologias de computação. O surgimento da Web na década de 1990 fez aumentar a importância e o interesse nessa área.

O baixo custo e a facilidade de aplicação a grandes repositórios de documentos incentivam o desenvolvimento de diversos métodos de indexação automática. Além disso, um sistema computacional irá realizar a indexação de maneira uniforme, utilizando sempre os mesmos critérios para o qual foi programado, independente da quantidade de documentos ou de qualquer fator externo.

Lancaster (2004, p.285) distingue dois tipos de indexação automática: *indexação por*

extração automática e indexação por atribuição automática. A indexação por extração automática é realizada geralmente por meio de cálculos matemáticos de frequência das palavras encontradas no texto de um documento. Na indexação por atribuição automática é utilizado um elemento externo ao texto, um vocabulário controlado, com o objetivo de normalizar os termos de indexação.

As ontologias, consideradas como vocabulários controlados, abrem novas perspectivas para as pesquisas em indexação automática, pois oferecem uma estrutura conceitual e terminológica restrita a um determinado domínio e originalmente representadas em linguagens legíveis por computador.

Nesse contexto, objetivamos delinear alguns princípios de utilização das características próprias das ontologias a fim de melhorar os atuais processos e métodos de indexação automática. Para tal, estabelecemos como objetivos específicos: realizar um levantamento bibliográfico sobre indexação automática e seus métodos e realizar estudos sobre as estruturas lógicas de construção de ontologias.

A presente pesquisa é de cunho teórico, caracterizada como **pesquisa bibliográfica** quanto aos meios ou à natureza das fontes utilizadas para abordagem e tratamento do objeto. Na tipologia de pesquisa estabelecida por Vergara (2004) definimos a presente pesquisa como **pesquisa exploratória** quanto ao seu fim.

2 VOCABULÁRIOS CONTROLADOS

O vocabulário controlado tem como objetivo: “reunir unidades depuradas de tudo aquilo que possa obscurecer o sentido: ambigüidade de vocábulo ou de construção, sinonímia, pobreza informativa, redundância, etc.” (CINTRA, 2002, p.42-43).

São três os tipos principais de vocabulários controlados utilizados na indexação manual: esquemas de classificação bibliográfica (como a Classificação Decimal de Dewey), listas de cabeçalhos de assuntos e tesouros, que procuram apresentar os termos tanto alfabética quanto sistematicamente. Nas classificações, o arranjo alfabético é secundário, na forma de um índice que remete para o arranjo principal, que é hierárquico. No tesouro, o arranjo explícito dos termos é alfabético, mas existe uma estrutura hierárquica implícita, incorporada à lista alfabética por meio de remissivas. A tradicional lista de cabeçalhos de assuntos, por sua vez, é similar ao tesouro por ser de base alfabética, mas difere dele porque incorpora uma estrutura hierárquica imperfeita e porque não distingue claramente as relações hierárquicas das relações associativas. Esses três tipos de vocabulário controlam sinônimos, distinguem homógrafos e agrupam termos afins, mas empregam métodos um tanto diferentes

para a linguagem controlada alcançar estes objetivos (LANCASTER, 2004).

O tesouro consolidou-se na Ciência da Informação como uma ferramenta bastante eficiente na representação da informação para recuperação, sendo largamente empregado ao longo dos anos por indexadores.

3 ONTOLOGIAS COMO VOCABULÁRIOS CONTROLADOS

A palavra ontologia é composta de duas outras: *onto* e *logia*. *Onto* deriva de dois substantivos gregos: *tà onta*, significando “o Ser”. A ontologia nasceu no seio da Filosofia, como o estudo ou conhecimento do Ser, dos entes e das coisas como elas são em si mesmas, remetendo-nos à metafísica de Aristóteles e suas categorias. Assim, a ontologia busca um sentido para a existência do ser, categorizando e classificando os seres para organizar e compreender o mundo (CHAUI, 2012).

A partir da década de 1990 o termo ontologia começa a ser frequentemente referenciado na área da Ciência da Computação, sobretudo na Inteligência Artificial. A Ciência da Computação, em especial a área de Inteligência Artificial, apropriou-se do termo ontologia. Assim a ontologia aparece em um novo contexto, um tanto distante da filosofia, agora como um mecanismo para organizar e categorizar conhecimento, representando e permitindo sua posterior recuperação.

A Ciência da Informação tem se preocupado com a organização do conhecimento desde sua origem, elaborando muitas ferramentas (como os vocabulários controlados) para que isso fosse feito de forma padronizada e consistente. Nesse contexto as ontologias aparecem como mais um mecanismo para essa finalidade, mas dessa vez com tecnologias e recursos legíveis por computador.

A definição clássica de ontologia no âmbito da Ciência da Computação é a de Gruber (1995), para o qual uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada. **Formal** diz respeito a “ser legível por computador”; **explícita**, indica que os elementos estão claramente definidos; **conceitualização** refere-se a um modelo abstrato de um fenômeno; e **compartilhada** significa que os conceitos presentes representam um conhecimento consensual, aceito por um grupo de pessoas.

Outra definição de ontologia bastante comum na literatura é a de Gómez-Pérez (1999), que afirma que uma ontologia consiste em um conjunto de termos ordenados hierarquicamente para descrever um domínio que pode ser usado como um esqueleto para uma base de conhecimentos.

Jasper e Uschold (1999) ressaltam que:

Uma ontologia pode possuir uma variedade de formas, mas necessariamente incluirá um vocabulário de termos, e alguma especificação de seus significados. Isto inclui definições e uma indicação de como conceitos estão inter-relacionados, o que impõe uma estrutura no domínio e restringe as possíveis interpretações dos termos.

As linguagens mais recentes de representação de ontologias são baseadas em Lógica descritiva, das quais destacamos a OWL, que atualmente é a linguagem recomendada pelo W3C para a construção de ontologias.

Uma ontologia pode ser desenvolvida em um editor de texto puro, escrevendo-se o código como se desenvolve um programa, mas o uso de ferramentas que auxiliem o processo de construção de ontologias contribui na agilidade de desenvolvimento, minimizando erros. (SANTARÉM SEGUNDO, 2010). Não há um padrão recomendado para construção de ontologias, mas existem diversas ferramentas disponíveis, tais como: OilEed, OntoEdit, Chimaera, API Jena, Protégé 2000, entre outras.

4 INDEXAÇÃO AUTOMÁTICA BASEADA EM ONTOLOGIAS

O início das pesquisas em indexação automática data da década de 1950, com os trabalhos de Hans Peter Luhn (SCHULTZ, 1968). De forma geral, os primeiros trabalhos nesse campo consideravam o texto de um documento como um elemento autônomo, cuja semântica se resolveria no interior do próprio texto. Posteriormente começaram a surgir pesquisas que utilizavam elementos externos aos documentos para dar suporte à indexação automática. Esses elementos podem ter diferentes níveis de complexidade, variando de simples listas de palavras até tesouros e ontologias.

Particularmente, as ontologias abrem novas perspectivas para as pesquisas em indexação automática, uma vez que oferecem uma estrutura conceitual e terminológica restrita a um determinado domínio e originalmente representadas em linguagens legíveis por computador, o que permite a sua utilização nos mais variados processos computacionais.

A utilização de uma base ontológica possibilita uma abordagem mais rica para a indexação, pois permite oferecer algum tipo de análise semântica. Essa análise pode ser efetuada a partir dos textos dos documentos, onde são identificados e selecionados termos que possam ser mapeados para os conceitos de uma determinada ontologia. Esse mapeamento permite padronizar o vocabulário e restringir o campo semântico dos termos, contextualizando-os ao domínio da ontologia, solucionando assim possíveis ambiguidades.

5 CONSIDERAÇÕES PARCIAIS

O uso de ontologia no processo de indexação permite agregar a esse processo não só uma terminologia de um domínio específico, mas também uma estrutura conceitual que pode ser utilizada para realizar inferências, e cujas relações permitem uma expansão dos termos extraídos por métodos puramente matemáticos.

Os vocabulários controlados, utilizados por indexadores para traduzir os assuntos de documentos em termos padronizados, sempre se mostraram eficientes para a representação da informação na indexação manual. A utilização das ontologias, vistas aqui como vocabulários controlados, trazem novas perspectivas para a indexação automática, visto que trazem novos recursos de relacionamentos semânticos e lógicos, não disponíveis nas tradicionais ferramentas utilizadas na biblioteconomia, tal como o tesauro.

REFERÊNCIAS

CHAUÍ, M. **Convite à Filosofia**. 14^a ed. São Paulo: Ática, 2012.

CINTRA, A. M. M. (Org.). **Para entender as linguagens documentárias**. 2.ed. São Paulo: Polis, 2002.

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

GÓMEZ-PÉREZ, A. Evaluation of taxonomic knowledge in ontologies and knowledge bases. *In: Twelfth Workshop on Knowledge Acquisition, Modeling and Management*, 12. Alberta, Canadá, 1999.

GRUBER, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal Human-Computer Studies** v.43, n.5-6, 1995.

GUARINO, N. Formal Ontology and Information Systems. *In: GUARINO, N. (ed.) Proceedings of FOIS'98*, Trento, Italy. Amsterdam: IOS Press, 1998.

JASPER, R.; USCHOLD, M. A Framework for understanding and classifying ontology applications. *In: KRR5-99*, Stockholm. 1999.

SANTAREM SEGUNDO, J E. **Representação Iterativa: um modelo para repositórios digitais**. 2010. 224f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

SCHULTZ, C. K. (ed.) **H.P. Luhn: Pioneer of information science: selected works**. New York: Spartan Books, 1968.

VERGARA, S. C. **Projetos e relatórios de pesquisa em Administração**. São Paulo: Atlas, 2004.