Um método de indexação automática baseada em ontologia

Maria Elisa Valentim Pickler Nicolino
Universidade Estadua Paulista – UNESP
Brasil · elisa@marilia.unesp.br

Eder Antonio Pansani Junior
Instituto Federal de Educação, Ciência e Tecnologia São Paulo
Brasil · epansani@ifsp.edu.br

Edberto Ferneda
Universidade Estadua Paulista – UNESP
Brasil · ferneda@marilia.unesp.br

Resumo: O processo de indexação tem como objetivo representar resumidamente o conteúdo informacional de documentos por meio de um conjunto de termos. Com o surgimento da Web, as pesquisas em indexação automática receberam grande impulso, tendo em vista a necessidade recuperação desse imenso acervo documental. As linguagens de indexação tradicionais, utilizadas para traduzir o conteúdo temático de documentos de forma padronizada, sempre se mostraram eficientes na indexação manual. As ontologias abrem novas perspectivas para as pesquisas em indexação automática, pois oferecem uma estrutura conceitual e terminológica restrita a um determinado domínio, representada com linguagens processáveis por computador. O uso de ontologias no processo de indexação automática permite agregar a esse processo uma terminologia de um domínio específico e uma estrutura lógica e conceitual que pode ser utilizada para realizar inferências, permitindo uma expansão dos termos diretamente extraídos do texto do documento. Este trabalho apresenta um método para a construção e a utilização de ontologias no processo de indexação automática. A partir dos resultados apresentados, pode-se conclui-se que a utilização de ontologias no processo de indexação permite não só agregar novos recursos ao processo de indexação, mas também permite pensar em novas e avançadas funcionalidades em um sistema de recuperação de informação.

Palavras-chave: Indexação Automática; Ontologia; Linguagem de indexação; SPQRQL.

Abstract: The indexing process aims to represent synthetically the informational content of documents by a set of terms whose meanings indicate the themes or subjects treated by them. With the emergence of the Web, research in automatic indexing received major boost with the necessity of retrieving documents from this huge collection. The traditional indexing languages, used to translate the thematic content of documents in standardized terms, always proved efficient in manual indexing. Ontologies open new perspectives for research in automatic indexing, offering a computer-processable language restricted to a particular domain. The use of ontologies in the automatic indexing process allows using a specific domain language and a logical and conceptual framework to make inferences, and whose relations allow an expansion of the terms extracted directly from the text of the document. This paper presents techniques for the construction and use of ontologies in the automatic indexing process. We conclude that the use of ontologies in the indexing process allows to add not only new feature to the indexing process, but also allows us to think in new and advanced features in an information retrieval system.

Keywords: Automatic Indexing; Ontology; Indexing Language; SPARQL.

1. Introdução

Indexar um documento visa representar o seu conteúdo informacional associando-lhe um conjunto de termos cujos significados remetem aos assuntos tratados por ele. A indexação tem por objetivo sintetizar um objeto linguístico, 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.

O processo de indexação realizado de forma manual, por seres humanos, é dependente de critérios subjetivos e pessoais, relacionados à formação e experiência do indexador. Assim, o tempo dispendido e a qualidade da indexação ficam atrelados a fatores não controláveis, o que acarreta uma variabilidade dos resultados de um indexador para outro, bem como de um mesmo em momentos diferentes.

Para Anderson e Perez-Carballo (2001), o baixo custo da indexação automática e sua facilidade de aplicação a grandes conjuntos de documentos incentivaram o desenvolvimento de métodos de indexação automática. Outra vantagem da indexação automatizada é a sua homogeneidade. Um sistema computacional irá realizar a indexação de maneira uniforme, utilizando sempre os mesmos critérios para o qual foi programado, independentemente de qualquer fator externo.

Os argumentos contra a indexação automatizada estão centrados na capacidade inerente do ser humano em tratar com a linguagem. Um indexador humano, utilizando o seu conhecimento, sua experiência e sua bagagem cultural pode reconhecer os diferentes significados de uma palavra ou frase em seus diferentes contextos. Tais significados, convertidos em novos termos de indexação, proporcionam uma melhor representação dos documentos e, consequentemente, melhoram a eficiência da recuperação de informação.

Os primeiros trabalhos em indexação automática consideravam o texto de um documento como um elemento autônomo, cuja semântica se resolveria no interior do próprio texto. Em abordagens posteriores começam a surgir pesquisas que utilizavam algum elemento externo aos documentos para dar suporte à indexação automática. Esses elementos podem ter diferentes níveis de complexidade, podendo variar de simples listas de palavras até tesauros e ontologias.

Particularmente, as ontologias abrem novas perspectivas para as pesquisas em indexação automática, pois oferecem uma estrutura conceitual e terminológica restrita a um determinado domínio, e originalmente representada em linguagens processáveis por computador.

Este trabalho apresenta um método de indexação automática no qual uma ontologia de domínio é vista como uma linguagem de indexação, utilizada para enriquecer a representação de documentos textuais agregando-lhes termos resultantes de inferências realizadas na ontologia.

2. Indexação

O processo de indexação visa representar um documento por meio de um conjunto de termos. Novellino (1996) afirma que:

> A principal característica do processo de representação da informação é a substituição de uma entidade linguística longa e complexa o texto do documento por sua descrição abreviada. O uso de tal sumarização não é apenas uma consequência de restrições práticas quanto ao volume de material a ser armazenado e recuperado. Essa sumarização é desejável, pois sua função é demonstrar a essência do documento. Ela funciona então como um artifício para enfatizar o que é essencial no documento considerando sua recuperação, sendo a solução ideal para organização e uso da informação (p.38).

A indexação caracteriza-se, portanto, como uma forma de representação de entidades linguísticas a fim de resumir o seu conteúdo e ressaltar a sua essência, permitindo ou facilitando a sua recuperação.

Restringindo-se a objetos textuais, Lancaster (2004, p.18) distingue dois tipos de indexação: *indexação por extração* e *indexação por atribuição*. Na indexação por

extração a seleção dos termos fica restrita ao contexto do próprio documento. O indexador, utilizando critérios institucionais e pessoais, seleciona no texto termos ou palavras que serão utilizados para representar o documento. Já a indexação por atribuição é realizada utilizando-se um elemento externo ao documento, um conjunto de termos previamente definidos e normalizados de complexidade variável. Após a leitura do texto, o indexador escolhe os termos mais adequados para representar o conteúdo informacional do documento.

O método de indexação proposto neste trabalho se caracteriza como um método de "indexação por atribuição automática" (Lancaster 2004, p.289) no qual as ontologias são consideradas e utilizadas como linguagens de indexação, com as quais é possível enriquecer a indexação de documentos textuais.

3. Ontologia

Segundo Soergel (1999) e Vickery (1997), o termo ontologia começou a ser utilizado na literatura da Ciência da Informação no final da década de 1990, principalmente por pesquisadores da área de Organização do Conhecimento. Nessa época, os instrumentos e métodos de classificação passaram a despertar um maior interesse de pesquisadores devido principalmente à necessidade de desenvolvimento de instrumentos de organização da informação no ambiente Web.

A Organização do Conhecimento vem se consolidando como um importante campo de investigação da Ciência da Informação a partir da fundação da *International Society for Knowledge Organization* (ISKO), em 1989, quando as principais ações para a consolidação da área foram adotadas.

Para Esteban Navarro (1996), a Organização do Conhecimento é a disciplina da Ciência da Informação que se dedica ao estudo dos fundamentos teóricos do tratamento e recuperação da informação, avaliando o uso de instrumentos lógico-linguísticos para controlar os processos de representação, classificação, ordenação e armazenamento do conteúdo informativo dos documentos com a finalidade de permitir sua recuperação e disseminação.

Uma definição clássica de ontologia no contexto da Ciência da Computação é a de Gruber (1995). Segundo esse autor, 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.

Gómez-Pérez (1999) 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. Uschold (1998) ressalta os componentes essenciais de uma ontologia:

> Uma ontologia pode assumir vários formatos, mas necessariamente deve incluir um vocabulário de termos e alguma especificação de seu significado. Esta deve abranger definições e uma indicação de como os conceitos estão inter-relacionados, o que resulta na estruturação do domínio e nas restrições de possíveis interpretações de seus termos.

Segundo Guizzard (2005), as ontologias consistem de uma estrutura formal de conceitos e relações de um domínio e um conjunto de axiomas que restringem a interpretação desta estrutura e permite a derivação de novos conhecimentos a partir do conhecimento real representado na estrutura de um determinado domínio

Segundo a W3C, uma ontologia é composta pela definição dos termos utilizados na descrição e na representação de uma área do conhecimento, e devem prover descrições para os seguintes tipos de conceitos:

- Classes – nos vários domínios de interesse;
- Relacionamentos entre essas classes (ou coisas);

- Propriedades (atributos) que essas classes (ou coisas) devem possuir.

Uma ontologia define os conceitos usados em uma determinada área de conhecimento, padronizando seus significados. Pode ser usada por pessoas, bases de dados e aplicações que precisam compartilhar informações e conceitos de um domínio (Daconta et al., 2003, p.167).

Ramalho (2010, p.38) apresenta resumidamente os componentes de uma ontologia:

- Classes e Subclasses: As classes e subclasses de uma ontologia agrupam um conjunto de elementos, "coisas", do "mundo real", que são representadas e categorizadas de acordo com suas similaridades, levando-se em consideração um domínio concreto. Os elementos podem representar coisas físicas ou conceituais, desde objetos inanimados até teorias científicas ou correntes teóricas;
- Propriedades Descritivas: Descrevem as características, adjetivos e/ou qualidades das classes;
- Propriedades Relacionais: Trata-se dos relacionamentos entre classes pertencentes ou não a uma mesma hierarquia, descrevendo e rotulando os tipos de relações existentes no domínio representado;
- Regras e Axiomas: Enunciados lógicos que possibilitam impor condições como tipos de valores aceitos, descrevendo formalmente as regras da ontologia e possibilitando a realização de inferências automáticas a partir de informações que não necessariamente foram explicitadas no domínio, mas que podem estar implícitas na estrutura da ontologia;
- Instâncias: Indicam os valores das classes e subclasses, constituindo uma representação de objetos ou indivíduos pertencentes ao domínio modelado, de acordo com as características das classes, relacionamentos e restrições definidas;
- Valores: Atribuem valores concretos às propriedades descritivas, indicando os formatos e tipos de valores aceitos em cada classe.

Guimarães (2002, p.53) apresentam algumas vantagens do uso de ontologias:

- Ontologias fornecem um vocabulário para representação do conhecimento. Vocabulário esse que traz uma conceitualização que o sustenta, evitando ambiguidades.
- Permitem o compartilhamento de conhecimento. Sendo assim, caso exista uma ontologia que modele adequadamente certo domínio do conhecimento, essa pode ser compartilhada e usada por pessoas que desenvolvam aplicações dentre desse mesmo domínio.
- Fornecem uma descrição exata do conhecimento. Diferentemente da Linguagem Natural, em que as palavras podem ter semântica diferente conforme o contexto, a ontologia é escrita em linguagem formal, ou seja, estabelecendo formalmente, então, as definições de um termo, eliminando ambiguidades.
- Há a possibilidade de mapeamento da linguagem da ontologia sem que com isso seja alterada a sua conceitualização, isto é, uma mesma conceitualização pode ser expressa em várias línguas.
- É possível estender o uso de uma ontologia genérica de forma que ela se adéque a um domínio específico.

A construção de uma ontologia pode ser pensada como uma união de peças que formam uma estrutura completa. Classes e subclasses definem um "esqueleto" na forma de uma hierarquia que pode ser expressa por meio de uma árvore ou de um grafo, complementada por propriedades descritivas, propriedades relacionais, regras e axiomas. A sua abrangência (domínio) deve ser previamente definida, e estabelece uma área do conhecimento ou uma parte do mundo que se pretende tratar.

Conforme o contexto apresentado, notamos que as ontologias se apresentam como um modelo de relacionamentos de entidades em um domínio particular do conhecimento. O objetivo principal de sua construção é a necessidade de um vocabulário compartilhado cujas informações possam ser trocadas e reusadas pelos seus usuários, sejam eles humanos ou agentes inteligentes (Santarem Segundo, 2010, p.104).

4. A Linguagem OWL

Web Ontology Language (OWL) é recomendada pelo consórcio W3C como a principal linguagem para a construção de ontologias. Essa linguagem tem como objetivo principal atender às necessidades de aplicação da Web Semântica e ser efetivamente utilizada por aplicações que necessitem processar o conteúdo de informações, e não somente apresentar a visualização destas informações. (Santarem Segundo, 2010, p.127).

A linguagem OWL foi projetada para prover uma linguagem que possa ser utilizada para descrever classes e seus relacionamentos em aplicações Web. Os elementos básicos para a construção de uma ontologia OWL são as classes, as instâncias das classes (indivíduos), propriedades e relacionamentos entre classes e instâncias.

A seguir são apresentados apenas alguns elementos da linguagem OWL considerados essenciais para o entendimento do método de indexação automática proposto neste trabalho.

4.1 Classes

Uma classe representa um grupo de indivíduos que compartilham algumas características ou propriedades comuns. Uma classe é utilizada para definir um conceito de um determinado domínio como, por exemplo, pessoas, automóveis, ou qualquer outra entidade concreta ou abstrata que se deseja representar. É importante observar que frequentemente a palavra *conceito* é utilizada como sinônimo de classe. Neste trabalho entende-se como *Classe* a representação concreta de um conceito.

Uma classe OWL é representada por meio da *tag* `owl:Class`, seguida de um atributo identificador (`rdf:ID`). É possível criar uma classe juntamente com algumas de suas características por meio da definição de um bloco delimitado pelas *tags* `<owl:Class>` e `</owl:Class>`.Entre o início e o final desse bloco é possível definir algumas propriedades e relações da classe que está sendo criada.

Toda ontologia deve se apoiar em uma estrutura taxionômica na qual as classes se organizam em uma forma hierárquica. Utilizando a linguagem OWL essa hierarquia de classes e subclasses pode ser criada utilizando a *tag* rdfs:`subClassOf`, como demonstrado na Figura 1.

As declarações apresentadas na Figura 1 mostram que as classes "Desktop" e "Notebook" são definidas como subclasses da classe "Computador". Essa construção estabelece relacionamentos de especialização e generalização, considerando o caminho que vai de uma classe para uma subclasse (especialização) ou de uma subclasse à sua correspondente classe superior (generalização).

Utilizando a OWL é possível definir duas classes como sendo equivalentes ou sinônimas. Feito isso, cada indivíduo de uma determinada classe é também membro da classe equivalente. A Figura 2 mostra a criação da classe "Laptop", definindo-a como equivalente à classe "Notebook".

Figura 1 – Hierarquia de classes

```
<owl:Class rdf:ID="Computador"/>
  <rdfs:label>Computador</rdfs:label>
...
<owl:Class rdf:ID="Desktop">
  <rdfs:label>Desktop</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Computador"/>
  </owl:Class>
...
<owl:Class rdf:ID="Notebook">
  <rdfs:label>Notebook</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Computador"/>
</owl:Class>
...
<owl:Class rdf:ID="AllInOne">
  <rdfs:label>All in One</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Desktop"/>
</owl:Class>
...
<owl:Class rdf:ID="Netbook">
  <rdfs:label>Netbook</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Notebook"/>
</owl:Class>
```

Fonte: elaborada pelos autores

Figura 2 – Classes equivalentes

```
<owl:Class rdf:ID="Laptop">
  <rdfs:label>Laptop</rdfs:label>

  <owl:equivalentClass rdf:resource="#Notebook"
</owl:Class>
```

Fonte: elaborada pelos autores

Essa funcionalidade pode ser utilizada para ligar conceitos de uma mesma ontologia, ou mesmo para efetuar a interoperabilidade entre duas ontologias diferentes.

4.2 Identificadores e Labels

As ontologias são utilizadas para representar o conhecimento sobre um determinado domínio por meio da descrição dos conceitos envolvidos nesse domínio. Esses conceitos são representados por meio de uma hierarquia de classes. Os elementos de uma ontologia são identificados por seus respectivos nomes cujos significados remetem à entidade que está sendo descrita.

Na linguagem OWL, o identificador (ID) de qualquer entidade não pode conter o caractere de espaço nem qualquer caractere especial. Assim, muitas vezes não é possível identificar uma determinada entidade da ontologia utilizando uma palavra ou um termo perfeitamente grafado, principalmente em idiomas que utilizam acentuação, tal como português, francês, alemão, espanhol etc.

Manaf et al. (2010) apresentam um estudo utilizando 306 ontologias OWL disponíveis na Web. A partir dessas ontologias identificou-se um conjunto de diferentes estilos de definição de identificadores de classes, indivíduos e propriedades dessas ontologias. Porém, as ontologias utilizadas na pesquisa tinham como base a língua inglesa, que possui certa facilidade na interpretação automática dos identificadores. Por exemplo, o identificador "AutomaticIndexing" (*camel case*) ou "Automatic_indexing" (*underscore*) podem ser facilmente interpretados como representantes do conceito "Automatic Indexing". Porém, em idiomas cujo léxico utiliza acentuação, essa interpretação será mais difícil.

Apesar dessas dificuldades, a linguagem OWL possui a propriedade *label* com a qual é possível a definição exata do termo que identifica uma entidade da ontologia utilizando caracteres de espaço e letras acentuadas. A utilização da propriedade *label*

é opcional, e qualquer elemento da ontologia pode possuir um ou mais *labels*. Essa propriedade permite também a especificação do idioma do termo com a utilização do parâmetro `xml:lang`. No exemplo da Figura 3, a classe denominada "IndexacaoAutomatica" é criada com a especificação em português e as traduções para as línguas espanhol, inglês e francês.

Como exposto, embora a linguagem OWL possua limitações quanto à forma como são identificados os elementos de uma ontologia, ela oferece recursos com os quais é possível apresentar tais identificadores de forma idêntica à linguagem natural. Por meio da propriedade *label* é possível não só descrever um determinado elemento de forma legível por humanos, mas também traduzir os identificadores em uma variedade de idiomas.

Figura 3 – Exemplo de Utilização da propriedade *label*

```
<owl:Class rdf:ID="IndexacaoAutomatica">
  <rdfs:label xml:lang="pt">Indexação Automática</rdfs:label>
  <rdfs:label xml:lang="sp">Indexación Automática</rdfs:label>
  <rdfs:label xml:lang="en">Automatic Indexing</rdfs:label>
  <rdfs:label xml:lang="fr">Indexation Automatique</rdfs:label>
…
</owl:Class>
```

Fonte: elaborada pelos autores

5. A linguagem SPARQL

SPARQL (*Protocol and RDF Language*) é recomendada pelo W3C como uma linguagem para expressar consultas e recuperar informações em diversas fontes de dados, desde que estejam armazenados nativamente em RDF (*Resource Description Framework*). Ramalho (2006) apresenta uma base histórica dos conceitos, propostas e tecnologias envolvidas no desenvolvimento da Web Semântica. Segundo o autor, as linguagens OWL e SPARQL começaram a figurar na arquitetura da Web Semântica no ano de 2005, como linguagem de desenvolvimento de ontologias e linguagem de consulta em estrutura RDF, respectivamente. (Ramalho, 2006, p.47).

De acordo com Pérez et al (2006), uma consulta SPARQL consiste em três partes:

- *Pattern Maching*: inclui várias características de combinação de padrões de gráficos, a união de padrões, aninhamento, filtragem e a possibilidade de escolher a fonte dos dados.
- *Solution modifiers*: permite modificar os valores de saída padrão aplicando alguns operadores como PROJECTION, LIMIT, DISTINCT, ORDER e OFFSET.
- *Output*: podem ser de diferentes tipos, como: booleanas, seleções, gráficos e descrições de recursos.

É importante destacar que estas partes não são obrigatórias para o funcionamento de toda e qualquer consulta, sendo facultado seu uso dependendo das necessidades de cada situação. Na Figura 4 pode ser visto um exemplo de uma consulta SPARQL simples.

Figura 3 - Exemplo de consulta SPARQL

```
1. # filename: ex003.rq
2.
3. PREFIX ab: <http://learningsparql.com/ns/addressbook#>
4.
5. SELECT ?craigEmail
6. WHERE
7. { ab:craig ab:email ?craigEmail . }
```

Fonte: DuCharme (2013, p.3)

Na Figura 4, na primeira linha o caractere "#" indica tratar-se de um comentário. Na terceira linha há a definição do prefixo "ab", que nomeia e resume a URL que aparece em seguida. Este prefixo será usado no código da consulta. Nas linhas de cinco a sete está a consulta em si. Neste exemplo a instrução usada é SELECT, indicando que o resultado será recuperado e apresentado em forma de uma seleção. O objeto de interesse é definido pela variável "?craigEmail", que aparece também na condição WHERE, indicando que o objetivo da consulta é recuperar o email de um recurso chamado "craig". Este exemplo utiliza o formato *Turtle triple*, que segue uma proposta mais simplista que os documentos no formato RDF/RDFS. No entanto, o uso de fontes de informação em RDF ou OWL pode ser feito apenas adicionando os respectivos prefixos e apontando a origem dos dados na execução da consulta.

Um processador SPARQL é um software capaz de rodar consultas em dados locais ou remotos. Para os testes realizados foi utilizado o *framework* Jena, um programa baseado em Java capaz de realizar consultas em ontologias em OWL utilizando a linguagem SPARQL.

Nos exemplos apresentados neste trabalho utilizaremos poucos recursos da linguagem SPARQL, visto que sua capacidade é muito maior que as necessidades apresentadas por esta pesquisa. A proposta é utilizar a instrução SELECT para retornar os resultados de uma busca por determinados relacionamentos contidos em uma ontologia OWL.

6. Método para a Utilização de Ontologias na Indexação Automática

A estrutura terminológica de uma ontologia é originalmente representada em linguagens processáveis por computador, o que permite sua utilização em vários processos computacionais, dentre eles a indexação automática.

Esta seção apresenta por meio de exemplos uma proposta de utilização de ontologias no processo de indexação automática. Será utilizada uma ontologia de termos de pediatria (PedTerm) que apresenta informações relacionadas à saúde e ao desenvolvimento infantil desde o pré-natal até os 21 anos de idade. Essa ontologia está disponível no BioPortal (http://bioportal.bioontology.org), um grande repositório de ontologias na área biomédica. Como foi originalmente criada no idioma Inglês, para cada classe incluímos propriedades *label* em três idiomas: inglês, espanhol e português.

6.1 Ontologias para indexação automática

A utilização de ontologias no processo de indexação se caracteriza como uma indexação por atribuição, na qual um único documento ou um conjunto de documentos (*corpus*) é vinculado a uma estrutura terminológica. Ao vincular um documento a uma determinada ontologia, declara-se indiretamente que os assuntos tratados pelos documentos estão relacionados ao domínio da ontologia. Isso permite restringir o campo semântico dos termos de indexação extraído dos documentos, minimizando ambiguidades.

A indexação por atribuição automática é realizada por meio da comparação entre termos extraídos dos textos de um *corpus* e um vocabulário do domínio. Portanto, é necessário existir uma coincidência entre os termos extraídos de um documento e os termos da ontologia. Porém, como visto, os identificadores das classes possuem a limitação de não permitir a utilização do caractere de espaço e de letras acentuadas. Tal limitação inviabiliza realizar comparações diretas entre termos extraídos dos textos e os identificadores dos elementos da ontologia. Assim, embora opcional, a utilização da propriedade *label* torna-se imprescindível na identificação dos elementos de uma ontologia para fins de indexação automática.

É possível ainda indicar o idioma do termo definido na propriedade *label* por meio do parâmetro `xml:lang`. Esse recurso permite o desenvolvimento de ontologias multilíngue, mesmo que os seus identificadores (IDs) sejam definidos em um determinado idioma. A Figura 5 apresenta as classes "Contaceptive_Device" e Condom" com as suas respectivas traduções definidas na propriedade *label*.

Para a proposta deste trabalho, a utilização da propriedade *label* é fundamental na criação de ontologias para fins de indexação. Portanto, a utilização das propriedades *label* deve ser obrigatória e os idiomas que serão utilizados nas suas traduções dependem do conhecimento do acervo documental a ser indexado.

Figura 4 – Classes com propriedades *label*

```
<owl:Class rdf:ID="Contraceptive_Device">
  <rdfs:label xml:lang="en">Contraceptive Device</rdfs:label>
  <rdfs:label xml:lang="es">Dispositivo
Anticonceptivos</rdfs:label>
  <rdfs:label xml:lang="pt">Dispositivo
Anticoncepcional</rdfs:label>

</owl:Class>

<owl:Class rdf:ID="Condom">
  <rdfs:label xml:lang="en">Condom</rdfs:label>
  <rdfs:label xml:lang="es">Condón</rdfs:label>
  <rdfs:label xml:lang="pt">Preservativo</rdfs:label>
  <rdfs:label xml:lang="pt">Camisinha</rdfs:label>

  <rdfs:subClassOf rdf:resource="#Contraceptive_Device"/>
</owl:Class>
```

Fonte: elaborada pelos autores com base na ontologia PedTerm

6.2 Extração de termos

A indexação por atribuição envolve, em um primeiro momento, uma indexação por extração, obtendo diretamente no texto do documento um conjunto de termos que serão utilizados para iniciar inferências na estrutura terminológica da linguagem de indexação utilizada. Esse processo de obter termos que indicam os assuntos tratados por um documento se estabeleceu como um campo de pesquisa na Ciência da Computação denominado "Extração de Informação" (*Information Extraction*) (SARAWAGI, 2008).

Extração de informação é a tarefa de extrair informação de forma automática a partir de documentos legíveis por computador. Essa extração pode ser realizada por meio de métodos puramente matemáticos (estatísticos) ou pela utilização de métodos e técnicas de Processamento de Linguagem Natural (Grishman, 1997).

Ao longo de mais de 50 anos de pesquisas em Indexação Automática, diversos métodos e algoritmos de extração de termos foram propostos e desenvolvidos. Desde os primeiros trabalhos de Luhn (Schultz, 1968), passando pelos trabalhos de Salton (Salton; Yang, 1973; Salton; McGill, 1983, p.131), até os métodos de indexação de páginas Web descritos por Keyser (2012, cap. 11). Diversos programas ou sistemas de extração de termos estão disponíveis gratuitamente na Web, não sendo objetivo deste trabalho apresentar tais métodos ou algoritmos. Assume-se a utilização de um sistema automatizado para a extração de um conjunto inicial de termos. A partir desses termos, o método aqui proposto irá agregar novos termos derivados de inferências em uma ontologia, buscando, assim, melhorar a representação dos documentos.

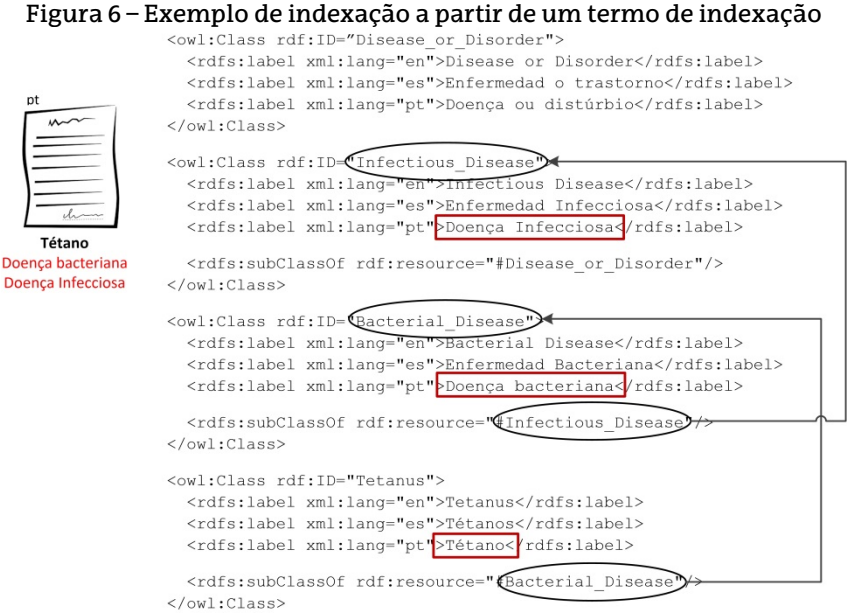
A fim de simplificar os exemplos apresentados a seguir, os documentos são inicialmente indexados por um único termo. Embora em um sistema real um documento possa ser indexado por um número variável de termos, o método aqui proposto seria aplicado a cada termo extraído do documento.

6.3 Atribuição de Conceitos

Um termo extraído do texto deve coincidir com um termo (conceito) definido na propriedade *label* de uma das classes da ontologia. Na ocorrência de tal coincidência, deve-se considerar o ID da classe à qual a propriedade *label* está associada para, a partir daí, realizar inferências ou traçar relacionamentos com outras classes da ontologia.

No exemplo da Figura 6 foi extraído do texto em português (pt) o termo "Tétano". Por meio de uma busca na ontologia encontrou-se esse termo na propriedade *label* pertencente à classe "Tetanus". Por meio da propriedade *subClassOf* verifica-se que esta classe é uma subclasse de "Bacterial_Disease". Segue-se, assim, para a classe "Bacterial_Disease" e obtém-se a propriedade *label* em português dessa classe: "Doença bacteriana". Repete-se o processo utilizando "Infectious_Disease" para acessar a classe de nível superior, obtendo-se o termo descrito na propriedade *label* em língua portuguesa "Doença infecciosa". Ao final desse processo o documento será representado pelos seguintes termos de indexação: "Tétano", "Doença bacteriana" e "Doença Infecciosa".

Figura 6 – Exemplo de indexação a partir de um termo de indexação



Fonte: elaborada pelos autores

Considerando um documento no formato OWL, como mostrado na Figura 6, foi criada a consulta SPARQL apresentada na Figura 7.

Figura 7 - Consulta SPARQL para execução do exemplo da Figura 6

```
PREFIX : <http://www.owl-ontologies.com/Ontology1358660052.owl>#
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT (str(?label) as ?Termos)
WHERE
{
 {
   ?res1 rdfs:label "Tétano"@pt .
   ?res1 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
  UNION
  {
   ?res1 rdfs:label "Tétano"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
  UNION
  {
   ?res1 rdfs:label "Tétano"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:subClassOf ?res3 .
   ?res3 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
}
```

Fonte: elaborada pelos autores

O número de termos de indexação atribuídos a um documento está relacionado principalmente às inferências realizadas nas classes mais genéricas. Em um sistema de indexação poderia ser definido um parâmetro numérico que definisse o número de classes que poderiam ser utilizadas no processo indexação de um documento. Esse parâmetro refletiria o nível de exaustividade da política de indexação. Por uma questão didática, para simplificar os exemplos, utilizamos apenas duas classes hierarquicamente superiores para indexar os documentos.

O idioma do documento a ser indexado deve ser conhecido para que as inferências na ontologia sejam realizadas nas propriedades *label* do idioma correspondente. No exemplo da Figura 6 e nos demais exemplos apresentados neste trabalho o idioma dos termos de indexação coincide com o idioma do documento. Porém, é possível realizar uma indexação cruzada (*cross-language indexing*), agregando termos de idiomas diferentes do idioma do documento.

6.4 Termos sinônimos

A propriedade *label* fornece uma maneira legível (por humanos) de descrever ou identificar uma classe. A OWL não impõe restrições quanto à sua utilização. Além de ser de uso opcional, é possível utilizar diversos *labels* em uma mesma classe. Pode também existir dois ou mais *labels* definidos com um mesmo valor no parâmetro *lang*.

No exemplo da Figura 8 foi extraído o termo "Varicela" de um documento em português (pt). Esse termo está presente em uma propriedades *label* em português da classe "Chiken_Pox". Porém, existe outra propriedade *label* em português contendo um sinônimo popular para essa doença ("Catapora") que poderá também fazer parte do índice do documento, juntamente com os termos relacionados às classes mais genéricas: "Doença Viral da Infância" e "Doenças ou Transtornos Pediátricos".

Figura 8 – Exemplo de atribuição de termos sinônimos

Fonte: elaborada pelos autores

Tomando como base o exemplo da Figura 8, foi elaborada a consulta mostrada na Figura 9, considerando a busca em uma ontologia OWL, que nestes exemplos é passado como um parâmetro para o processador Jena.

A consulta da Figura 9 é muito semelhante à consulta da Figura 7, sendo que a diferença recai apenas no termo informado.

Outra possibilidade é considerar como sinônimas todas as classes equivalentes associadas às classes referenciadas durante o processo de indexação. No exemplo da Figura 10, o termo "Espinha" é extraído do documento em português, que está

representado na propriedade *label* (pt) da classe "Pimple" da ontologia. A classe "Pimple" possui uma classe equivalente (*equivalantClass*), identificada por "Acne". O sistema segue assim para a classe "Acne" e atribui ao documento o valor de *label* em português.

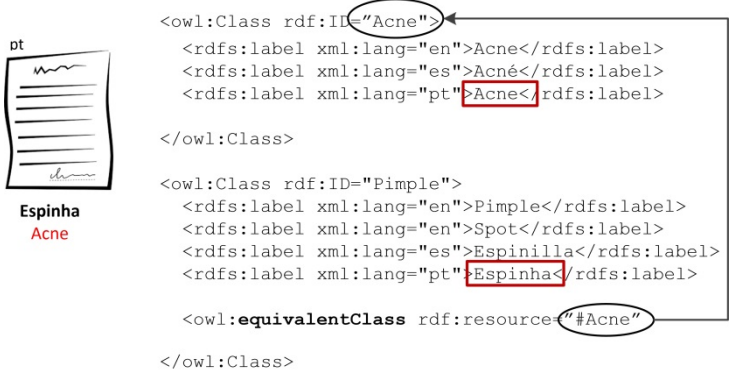
Figura 9 - Consulta SPARQL para execução do exemplo da Figura 8

```
PREFIX : <http://www.owl-ontologies.com/Ontology1358660052.owl>#
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT (str(?label) as ?Termos)
WHERE
{
  {
   ?res1 rdfs:label "Varicela"@pt .
   ?res1 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
  UNION
  {
   ?res1 rdfs:label "Varicela"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
  UNION
  {
   ?res1 rdfs:label "Varicela"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:subClassOf ?res3 .
   ?res3 rdfs:label ?label
   FILTER(lang(?label) = "pt") .
  }
}
```

Fonte: elaborada pelos autores

Figura 10 – Utilização de classes equivalentes como termos sinônimos



Fonte: elaborada pelos autores

Uma ontologia pode possuir diversos tipos e uma grande quantidade de relacionamentos. A adequada utilização desses relacionamentos em um sistema automatizado poderá resultar uma indexação mais eficiente.

Indexação Multilíngue

A propriedade *label* possui o parâmetro `xml:lang`, que permite a especificação (tradução) de identificadores da ontologia em vários idiomas. Com isso, uma mesma ontologia pode ser utilizada na indexação de um *corpus* contendo documentos de diferentes idiomas.

O exemplo da Figura 11 apresenta um *corpus* contendo três documentos de idiomas diferentes: inglês ("en"), espanhol ("es") e português ("pt"), e uma ontologia cujos identificadores de classes estão traduzidos para esses três idiomas. O processo de

indexação automática se dará de forma semelhante aos apresentados nos exemplos anteriores.

Figura 11 – Indexação de um *corpus* multilíngue

Fonte: elaborada pelos autores

A Figura 12 apresenta a consulta SPARQL elaborada para o exemplo da Figura 10, onde propõe-se a indexação multilíngue. O detalhe mais significativo neste código é a supressão do modificador FILTER, era responsável por filtrar e retornar apenas os termos em língua portuguesa. Neste exemplo o termo informado "Hipotireoidismo" vem seguido da instrução "@pt" para especificar que ele está grafado neste idioma, no entanto o resultado é um conjunto de termos de todos os idiomas disponíveis na ontologia, no caso português, inglês e espanhol.

Figura 12 - Consulta SPARQL para execução do exemplo da Figura 10

```
PREFIX : <http://www.owl-ontologies.com/Ontology1358660052.owl>#
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT (str(?label) as ?Termos)
WHERE
{
  {
   ?res1 rdfs:label "Hipotireoidismo"@pt .
   ?res1 rdfs:label ?label
  }
  UNION
  {
   ?res1 rdfs:label "Hipotireoidismo"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:label ?label
  }
  UNION
  {
   ?res1 rdfs:label "Hipotireoidismo"@pt .
   ?res1 rdfs:subClassOf ?res2 .
   ?res2 rdfs:subClassOf ?res3 .
   ?res3 rdfs:label ?label
  }
}
```

Fonte: elaborada pelos autores

É possível, assim, desenvolver sistemas automatizados para a indexação de uma grande quantidade de documentos de diferentes idiomas utilizando uma única ontologia adequadamente construída.

7. Conclusão

Por meio de exemplos implementados na linguagem SPARQL, este trabalho propôs um método de utilização de ontologias no processo de indexação automática.

Uma ontologia possui necessariamente um vocabulário de termos restritos a um domínio. Nesse trabalho, considerou-se o vocabulário de domínio presente em toda e qualquer ontologia como uma linguagem de indexação capaz auxiliar em processo de indexação automática por atribuição.

As restrições impostas pela linguagem OWL na formação dos identificadores dos elementos de uma ontologia impõem a utilização de recursos geralmente negligenciados na criação de ontologias, como é o caso da propriedade *label*. Portanto, a utilização de ontologias no processo de indexação automática parte do desenvolvimento de ontologias direcionadas para essa finalidade.

Por meio dos exemplos apresentados é possível verificar a exequibilidade e o potencial de sistemas automatizados de indexação baseados em ontologias. Como visto, é possível utilizar uma mesma ontologia para indexar documentos de diferentes idiomas, o que permite o desenvolvimento de sistemas de recuperação de informação conhecidos como *cross language*. Outra possibilidade é a utilização de uma ontologia como elemento principal de uma ferramenta de busca, na qual os documentos e as buscas são representados a partir de uma mesma ontologia. A partir de uma ontologia, representada de forma gráfica e dinâmica, o usuário poderia ter acesso à terminologia de uma área do conhecimento, no idioma que desejasse. A especificação da busca seria realizada pela seleção dos termos na interface.

Enfim, acredita-se que o método exposto neste trabalho pode vir a ser utilizado em diversas ideias adjacentes, sendo possível imaginar funcionalidades que podem ser desenvolvidas sistemas de indexação automática ou sistemas de recuperação de informação.

Referências

Anderson, J.D., Perez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing and Management*, 37, p.231-254.

Daconta, M.C., Obrst, L.J., Smith, K.T. (2003). *The Semantic Web*: a guide to the Future of XML, Web Services, and Knowledge Management. Indianápolis: Wiley Publishing.

Ducharme, Bob. (2013). *Learning SPARQL*: Querying and Updating with SPARQL 1.1. 2nd ed. Sebastopol (USA): O' Reilly.

Esteban Navarro, M.A. (1996). El marco disciplinar de los lenguajes documentales: la Organización del Conocimiento y las ciencias sociales. **Scire**, Zaragoza, 2(1).

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

Grishman, Ralph. (1997). Information extraction, techniques and challenges. In: *International Summer School* SCIE-97. New York. Proceedings... New York : Springer-Verlag.

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

Guimarães, F.J.Z. (2002). *Utilização de ontologias no domínio B2C*. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Guizzard G. (2005). *Ontological Foundations for Structural Concept Models*, Ph.D. Thesis, University of Twente, The Netherlands.

Keyser, P. (2012). *Indexing: from thesauri to the Semantic Web*. Burlington, MA: Elsevier Science.

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

Manaf, N.A.A., Bechhofer, S. & Stevens, R. (2010). A Survey of Identiers and Labels in OWL Ontologies. *Proceedings of the 6th International Workshop on OWL Experiences and Directions* (OWLED).

Novellino, M.S.F. (1996). Instrumentos e metodologias de representação da informação. *Informação & Informação*, Londrina, 1(2), p.37-45.

Pérez, J., Arenas, M., & Gutierrez, C. (2006). *Semantics and complexity of SPARQL*. The Semantic Web-ISWC 2006, p.30–43.

Ramalho, R.A.S. (2010). *Desenvolvimento e utilização de ontologias em Bibliotecas Digitais*: uma proposta de aplicação. Marília, 2010. 145 f. Tese (Doutorado em Ciência da Informação). Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Campus de Marília.

Ramalho, R.A.S. (2006). *Web Semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação*. Dissertação (Mestrado em Ciência da Informação). Faculdade de Filosofia e Ciências – Universidade Estadual Paulista, Marília.

Salton, G., & Yang, C.S. (1973). On the especification of term values in automatic indexing. *Journal of the Americam Society for Information Science*, 26(1).

Salton, G.,& McGill, J.M. (1983). *Introduction to Modern Information Retrieval*. New York, McGraw-Hill.

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

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases*, 1(3).

Schultz, C.K. (Ed.). (1968). *H. P. Luhn: Pioneer of information science: Selected works*. New York: Spartan Books.

Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*. 50(12).

Uschold, M. (1998) Knowledge level modelling concepts and terminology. *The Knowledge Engineering Review*, 13(1), p.5-29.

Vickery, B. C. (1997). Ontologies. *Journal of Information Science*. 23(4).