Um Método para a Utilização de Ontologias na Indexação Automática

Maria Elisa Valentim Pickler Nicolino
Universidade Estadual Paulista (Unesp), Brasil. E-mail: elisa@marilia.unesp.br

Edberto Ferneda
Universidade Estadual Paulista (Unesp), Brasil. E-mail: ferneda@marilia.unesp.br

Resumo
O processo de indexação tem como objetivo representar sinteticamente o conteúdo informacional de documentos por meio de um conjunto de termos cujos significados remetem aos temas ou assuntos tratados por eles. 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 em termos padronizados 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 e originalmente representada em linguagens processáveis por computador. O uso de ontologias no processo de indexação automática permite agregar a esse processo uma linguagem 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 diretrizes técnicas para a construção e utilização de ontologias no processo de indexação automática por meio de exemplos. 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.

1 Introdução

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

Embora a prática da indexação possa ser regulada por políticas e princípios institucionais, a eficiência do processo de indexação manual, realizado por seres humanos, é dependente de critérios subjetivos e pessoais, relacionados à formação e experiência do indexador. Assim, o tempo despendido e a qualidade da indexação manual ficam atrelados a fatores não controláveis, o que pode afetar no custo desse processo.

As dificuldades inerentes à indexação manual e a grande quantidade de documentos publicados e disponibilizados, que caracterizou a "explosão informacional", justificaram estudos que buscavam formas de auxiliar o indexador no exercício de sua atividade. As primeiras pesquisas em indexação automática aconteceram no final dos anos de 1950, época de rápido desenvolvimento das tecnologias de computação. O surgimento da Web nos anos de 1990 deu um grande

impulso nas pesquisas sobre o tema, tendo o vista a necessidade de se operar esse imenso acervo documental.

Anderson e Perez-Carballo (2001) citam o baixo custo da indexação automática e sua facilidade de aplicação a grandes conjuntos de documentos como fatores que incentivaram o desenvolvimento de métodos de indexação automática. Outro argumento em favor da indexação automatizada está na homogeneidade desse processo quando realizados por algoritmos computacionais. O resultado da indexação manual (intelectual) pode variar de um indexador para outro, bem como de um mesmo indexador em momentos diferentes. Um sistema computacional irá realizar a indexação de maneira uniforme, utilizando sempre os mesmos critérios para o qual foi programado, independentemente da quantidade de documentos ou de qualquer fator externo.

Os argumentos contra a indexação automatizada estão centrados na capacidade inerente do ser humano em tratar com a linguagem. Para um ser humano as palavras deixam de ser meros dados vazios de significado e tornam-se formas de representação mental de elementos do conhecimento. Assim, um indexador humano, utilizando o seu conhecimento 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 melhoria na representação dos documentos de um *corpus*, melhorando, por conseguinte, a eficiência do processo de recuperação de informação.

Os primeiros trabalhos nesse campo 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 legíveis 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 indexação de documentos textuais agregando-lhes novos termos, derivados de inferências realizadas sobre a ontologia. Embora tal proposta não tenha partido de resultados práticos ou experimentais, baseiam-se em pesquisa bibliográfica na área de Ciência da Computação, em trabalhos que relatam o funcionamento de alguns sistemas já desenvolvidos ou em desenvolvimento.

2 Indexação

Pinto (2010) afirma que a representação das coisas está atrelada ao conceito de substituição, já que quando representamos criamos uma relação entre o que se apresenta e o signo, em um ato de substituição, mas que não pode deixar de ser um ato de conhecimento. Para o autor, conhecer é classificar, porque para classificar é preciso possuir os fundamentos dos conceitos que serão classificados, de forma que, ao ser constituído, o signo propicia um conhecer, justamente porque se coloca como representante de conceitos, num jogo de substituição.

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.

A indexação caracteriza-se, portanto, como uma forma de representação de entidades linguística a fim de sumarizar 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 (léxico) cuja complexidade pode variar deste uma lista de cabeçalhos de assunto, um tesauro ou uma ontologia. Após a leitura do texto, o indexador escolhe os termos mais adequados para representar o conteúdo informacional do documento.

Neste trabalho as ontologias são consideradas e utilizadas como linguagens de indexação, com as quais é possível enriquecer a indexação de documentos textuais.

2.1 Linguagens de Indexação

Na literatura da área da Ciência da Informação sobre indexação encontramos os termos "linguagens de indexação", "linguagens documentárias" e "vocabulários controlados" muitas vezes usados como sinônimos; outras vezes são utilizados de forma diferenciada, mas geralmente tais diferenças não são criteriosamente apresentadas. Neste trabalho utiliza-se preferencialmente o termo "linguagem de indexação" para referenciar qualquer estrutura terminológica utilizada no processo de indexação por atribuição. No entanto será respeitada e conservada a denominação utilizada por cada autor.

Lancaster (2004, p.19) utiliza o termo "vocabulário controlado" e o define como "uma lista de termos autorizados" que o indexador poderá atribuir a um documento. Esses termos servirão como pontos de acesso mediante os quais um item será localizado e recuperado no momento da busca por um documento. Ainda segundo o autor, um vocabulário controlado costuma ser mais do que uma simples lista de termos, pois inclui, em geral, uma forma de estrutura semântica que se destina especialmente a:

- Controlar sinônimos, optando por uma forma única e padronizada com remissiva de todas as outras;
- Diferenciar homógrafos;
- Reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si.

Na Ciência da Informação, o tesauro consolidou-se como uma ferramenta bastante eficiente na representação da informação para fins de indexação e recuperação, sendo largamente empregado por indexadores e demais profissional da informação.

Um tesauro é composto de um conjunto de descritores ordenados segundo as relações recíprocas existentes entre eles. Estudos como os de Pickler (2007), Feitosa (2006), Sales e Café (2008), entre outros, evidenciem semelhanças e diferenças entre tesauros e ontologias, uma vez que ambos procuram representar o conhecimento para sua posterior recuperação.

> Do ponto de vista da representação do conhecimento, uma ontologia não deve ser concebida apenas como um vocabulário informal, ou mesmo como

> uma linguagem de termos estruturados – como um tesauro, por exemplo -, mas requer uma possibilidade de interpretação algorítmica dos seus significados e, por conseguinte, uma representação em uma linguagem formal, cujo processamento dos significados pode ser realizado por máquinas. Dito de outro modo: uma ontologia requer a explicitação lógico-formal de significados e palavras, que devem ser expressos por meio de construtos matemáticos (FEITOSA, 2006, p. 73).

Embora os tesauros sejam ampla e tradicionalmente utilizados na indexação manual, as ontologias se apresentam como uma nova tecnologia para auxiliar na organização da informação, especificamente no processo de indexação automática, como propomos no presente trabalho.

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 da comunidade de Ciência da Computação, 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.

Segundo Ramalho (2010, p.37):

> Entre os instrumentos de representação tradicionalmente utilizados na área de Ciência da Informação, os tesauros apresentam-se como os que possuem maior aproximação com as ontologias, devido ao fato de ambos os instrumentos serem constituídos por meio de linguagens de estruturas combinatórias, de caráter especializado, representando termos e conceitos organizados a partir de tipos de relacionamentos.
>
> Ao longo dos últimos anos inúmeros estudos comparativos entre ontologias e tesauros têm constatado que, apesar de possuírem características comuns, tais instrumentos caracterizam-se como diferentes modelos de representação do conhecimento. Enquanto os tesauros são desenvolvidos como ferramentas de auxílio para os usuários na busca de informações, as ontologias têm como principal objetivo descrever formalmente os recursos informacionais para possibilitar a realização de inferências automáticas.

Uma 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. Breitman (2005, p.31) nos apresenta a definição de Uschold e Jasper para o termo:

> 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.

Uma ontologia é composta de um conjunto de conceitos dentro de um determinado domínio, organizados em uma taxonomia. Taxonomia é a classificação de informações no formato de uma hierarquia, de acordo com os relacionamentos que estabelecem com entidades do mundo real que representam. Assim, servem justamente para classificar informação em uma hierarquia, utilizando o relacionamento de generalização ("tipo-de" ou "pai-filho"). Dessa forma, em uma taxonomia a generalização/especialização é o único tipo de relacionamento que existe entre seus termos (BREITMAN, 2005).

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.

Guimarães (2002, p.53) apresentam algumas vantagens do uso de ontologias na Ciência da Computação. São eles:

- 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.

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; OBRST, SMITH, 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.

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 apresentam-se 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

Atualmente a 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 de ontologia que possa ser usada para descrever, de uma forma natural, classes e relacionamentos em documentos e 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.

Neste trabalho serão apresentados apenas alguns elementos da linguagem OWL considerados essenciais para o entendimento do método aqui proposto. Em pesquisas futuras será realizado de um

estudo extensivo e detalhado dessa linguagem a fim de aperfeiçoar este método.

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 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 do bloco OWL é possível definir algumas propriedades e relações da classe que está sendo criada, reduzindo o código e facilitando a leitura e a compreensão da estrutura da ontologia.

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 no Exemplo 1.

Exemplo 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>
```

As declarações apresentadas no Exemplo 1 mostram que as classes "Desktop" e "Notebook" são definidas como subclasses da classe "Computador". Essa construção estabelece relacionamentos de especificidade e generalização, considerando o caminho que vai de uma classe para uma subclasse (especificidade) ou de uma subclasse à sua correspondente classe superior (generalização). Na Ciência da Computação é comum denominar os relacionamentos entre classes e subclasses como "tipo-de" (*type-of*) ou "é-um" (*is-a*). Assim, por exemplo, diz-se que Notebook *é-um* ou é um *tipo-de* Computador. Da mesma forma, Desktop é também um *tipo-de* Computador.

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. O Exemplo 2 mostra a criação da classe "Laptop" definindo-a como equivalente à classe "Notebook".

Exemplo 2 – Classes equivalentes

```
<owl:Class rdf:ID="Laptop">
  <rdfs:label>Laptop</rdfs:label>
  <owl:equivalentClass rdf:resource="#Notebook"
</owl:Class>
```

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 e relacionamentos envolvidos nesse domínio. Esses conceitos são representados por classes e indivíduos, e os relacionamentos são definidos por meio de propriedades. Classes, indivíduos e propriedades 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 e nem caracteres especiais, incluindo aí as letras acentuadas e o "ç". 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, espanhol etc.

Manaf, Bechhofer e Stevens (2010) apresentam um estudo utilizando um *corpus* contendo 306 ontologias OWL disponíveis na Web. A partir desse *corpus* identificou-se um conjunto de diferentes estilos utilizados na definição de identificadores de classes, indivíduos e propriedades dessas ontologias (Tabela 1).

Tabela 1 – Estilos de Identificadores em Ontologias

Estilo	Exemplo
Camel Case	`IndexacaoBaseadaEmOntologia`
Underscore	`Indexacao_baseada_em_ontologia`
Hífen	`Indexacao-baseada-em-ontologia`
Camel Case + Underscore	`IndexacaoBaseada_em_ontologia`
Camel Case + Hífen	`IndexacaoBaseada-em-ontologia`
Hífen + Underscore	`Indexacao_Baseada-em-ontologia`
Palavra única	`Indexacao`

Segundo o referido trabalho, a maioria das ontologias estudadas utilizavam identificadores definidos em algum desses estilos. Porém, há de se considerar que 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.

A linguagem OWL possui a propriedade *label* com a qual é possível a definição exata do termo que identifica uma entidade (classe, indivíduo ou propriedade), possibilitando a utilização de caracteres de espaço e letras acentuadas. A forma mais simples de utilização da propriedade *label* é apresentada no Exemplo 3.

Exemplo 3 – Forma simplificada de utilização da propriedade *label*

```
<owl:Class rdf:ID="IndexacaoAutomatica">
  <rdfs:label> Indexação Automática </rdfs:label>
...
</owl:Class>
```

A utilização da propriedade *label* é opcional e qualquer entidade, seja classe, indivíduo ou propriedade, pode possuir uma ou mais *labels*. Essa propriedade permite também a especificação do idioma do termo, com a utilização do parâmetro `xml:lang`. No Exemplo 4, a classe denominada "IndexacaoAutomatica" é criada com a especificação (`rdfs:label`) em português e as traduções para as línguas espanhol, inglês e francês.

Exemplo 4 – 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">L'indexation automatique</rdfs:label>
…
</owl:Class>
```

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 para humanos, mas também traduzir os identificadores em uma variedade de idiomas.

5 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[1], 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.

5.1 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.

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. O Exemplo 5 apresenta as classes “Contaceptive_Device” e “Condom” com as suas respectivas traduções definidas na propriedade *label*.

[1]http://bioportal.bioontology.org/

Exemplo 5 – 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>
```

Para a proposta deste trabalho, a utilização da propriedade *label* é de importância 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.

5.2 Extração de termos

A indexação por atribuição envolve, em um primeiro momento, uma indexação por extração, obtendo diretamente no documento um conjunto de termos que serão utilizados para iniciar inferências na estrutura terminológica 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 que serão utilizados para representar o conteúdo informacional dos documentos. 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, a forma de funcionamento do método aqui proposto seria aplicado a cada termo extraído do documento.

5.3 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 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".

Exemplo 6–Indexação automática a partir de um termo de indexação

Tétano
Doença bacteriana
Doença Infecciosa

```
<owl:Class rdf:ID=”Disease_or_Disorder">
  <rdfs:label xml:lang="en">Disease or Disorder</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedad o trastorno</rdfs:label>
  <rdfs:label xml:lang="pt">Doença ou distúrbio</rdfs:label>
</owl:Class>

<owl:Class rdf:ID="Infectious_Disease">
  <rdfs:label xml:lang="en">Infectious Disease</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedad Infecciosa</rdfs:label>
  <rdfs:label xml:lang="pt">Doença Infecciosa</rdfs:label>

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

<owl:Class rdf:ID="Bacterial_Disease">
  <rdfs:label xml:lang="en">Bacterial Disease</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedad Bacteriana</rdfs:label>
  <rdfs:label xml:lang="pt">Doença bacteriana</rdfs:label>

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

<owl:Class rdf:ID="Tetanus">
  <rdfs:label xml:lang="en">Tetanus</rdfs:label>
  <rdfs:label xml:lang="es">Tétanos</rdfs:label>
  <rdfs:label xml:lang="pt">Tétano</rdfs:label>

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

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 mais genéricas que poderiam ser utilizadas no processo indexação de um documento. Esse parâmetro refletiria o nível de exaustividade e especificidade 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 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.

5.4 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 7 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".

Exemplo 7 – Atribuição de termos sinônimos


```
<owl:Class rdf:ID="Disease_or_Disorder">
  <rdfs:label xml:lang="en">Disease or Disorder</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedad o trastorno</rdfs:label>
  <rdfs:label xml:lang="pt">Doença ou distúrbio</rdfs:label>
</owl:Class>

<owl:Class rdf:ID="Pediatric_Disease_or_Disorder">
  <rdfs:label xml:lang="en">Pediatric Disease or Disorder</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedades o Trastornos Pediátricos</rdfs:label>
  <rdfs:label xml:lang="pt">Doenças ou Transtornos Pediátricos</rdfs:label>

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

<owl:Class rdf:ID="Childhood_Viral_Disease">
  <rdfs:label xml:lang="en">Childhood Viral Disease</rdfs:label>
  <rdfs:label xml:lang="es">Enfermedad viral de la infancia</rdfs:label>
  <rdfs:label xml:lang="pt">Doença Viral da infância</rdfs:label>

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

<owl:Class rdf:ID="Chicken_Pox">
      <rdfs:label xml:lang="en">Chicken pox</rdfs:label>
      <rdfs:label xml:lang="es">Varicela</rdfs:label>
      <rdfs:label xml:lang="pt">Varicela</rdfs:label>
      <rdfs:label xml:lang="pt">Catapora</rdfs:label>

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

Outra possibilidade é considerar como sinônimas todas as classes equivalentes associadas às classes referenciadas durante o processo de indexação. No Exemplo 8 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.

Exemplo 8 – Utilização de classes equivalentes como termos sinônimos


```
<owl:Class rdf:ID="Acne">
  <rdfs:label xml:lang="en">Acne</rdfs:label>
  <rdfs:label xml:lang="es">Acné</rdfs:label>
  <rdfs:label xml:lang="pt">Acne</rdfs:label>

</owl:Class>

<owl:Class rdf:ID="Pimple">
  <rdfs:label xml:lang="en">Pimple</rdfs:label>
  <rdfs:label xml:lang="en">Spot</rdfs:label>
  <rdfs:label xml:lang="es">Espinilla</rdfs:label>
  <rdfs:label xml:lang="pt">Espinha</rdfs:label>

  <owl:equivalentClass rdf:resource="#Acne"

</owl:Class>
```

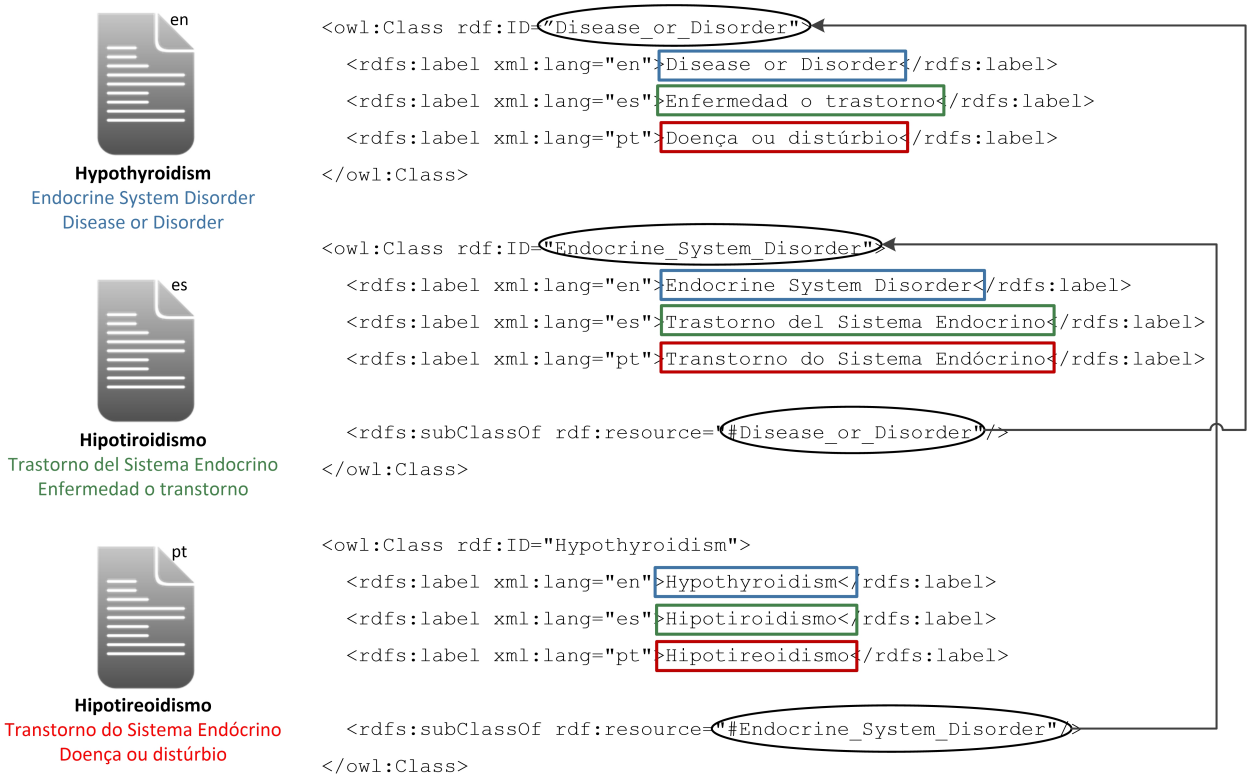
Uma ontologia pode possuir diversos tipos e uma grande quantidade de relacionamentos. A adequada utilização desses relacionamentos em um sistema automatizado poderá resultar uma indexação mais eficiente.

5.5 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 9 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.

Exemplo 9 – Indexação de um *corpus* multilíngue



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

6 Conclusões

Por meio de exemplos, este trabalho propôs um método de utilização de ontologias no processo de indexação automática. Embora tal método não seja resultante de um trabalho experimental, ele deriva de estudos de sistemas baseados em ontologia apresentados em diversos tipos de trabalhos (artigos, teses, dissertações) disponíveis na Web.

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, utilizando como elemento principal a identificação (ID) das classes da ontologia.

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 em 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 diversas funcionalidades que podem ser desenvolvidas em um sistema de recuperação de informação baseado em ontologia. Atualmente está sendo desenvolvido um protótipo computacional que irá demonstrar e validar as ideias e o método aqui apresentado.

A Method for Using Ontologies on Automatic Indexing

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-process able 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.*

Referências

ANDERSON, J. D.; PEREZ-CARBALLO, J. 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**, v.37, p.231-254. 2001.

ABNT - ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: métodos para análise de documentos: determinação de seus assuntos e seleção de termos de indexação: procedimento. Rio de Janeiro, 1992.

BREITMAN, Karin. **Web Semântica:** a internet do futuro. Rio de Janeiro: LTC, 2005.

CAVALCANTI, Cordélia R. **Indexação e tesauro**: metodologia e técnicas. Brasília: Associação de Bibliotecários do Distrito Federal, 1978.

DACONTA, M.C.; OBRST, L.J.; SMITH, K.T.**The Semantic Web**: a guide to the Future of XML, Web Services, and Knowledge Management. Indianápolis: Wiley Publishing, 2003.

DAHLBERG, Ingetraut. A referent-oriented, analytical concept theory for Interconcept. **International Classification**, Frankfurt, v.5, n.3, 1978.

ESTEBAN NAVARRO, M.A. El marco disciplinar de los lenguajes documentales: la Organización del Conocimiento y las ciencias sociales. **Scire**, Zaragoza, v.2, n.1, 1996.

EUZENAT, J.; SHVAIKO, P. **Ontology Matching**. 2.ed. Springer-Verlag, 2007.

FEITOSA, A. **Organização da informação na web: das *tags* à web** semântica. Brasília: Thesaurus, 2006.

FUJITA, M.S.L., Avaliação da eficácia de recuperação do sistema de indexação PRECIS. **Ciência da Informação**, v. 18, n.2, 1989.

FUJITA, M.S.L. Organização e representação do conhecimento no Brasil: análise de aspectos conceituais e da produção científica do ENANCIB no período de 2005 a 2007. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v.1, n.1, 2008.

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.

GRISHMAN, Ralph. Information extraction; techniques and challenges. In: **International Summer School** SCIE-97, 1997, New York. Proceedings… New York : Springer-Verlag, 1997.

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

GUIMARÃES, F. J. Z. **Ontologies use in B2C domain**, 2002. 195p. Dissertação (Mestrado em Informática) - Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

KEYSER, P. **Indexing**: from thesauri to the Semantic Web. Burlington, MA: Elsevier Science, 2012. 273 p

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

LOPES, I.L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**: Brasília, 2002, v. 31, n. 1, p. 41-52.

MANAF, Nor Azlinayati Abdul; BECHHOFER, Sean; STEVENS, Robert. A Survey of Identiers and Labels in OWL Ontologies. **Proceedings of the 6th International Workshop on OWL Experiences and Directions (OWLED)**, 2010.

NOVELLINO, Maria Salet Ferreira. Instrumentos e metodologias de representação da informação. **Informação & Informação**, Londrina, v.1, n.2, p.37-45, jul./dez. 1996.

PICKLER, Maria Elisa Valentim. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em Ciência da Informação**, v.12, n.1, p.65-83, abr. 2007

PINTO, Lourival Pereira. A recepção da informação: apresentação ou representação? **DataGramaZero** - Revista de Ciência da Informação, v.11, n.5, 2010.

RAMALHO, R.A.S. **Desenvolvimento e utilização de ontologias em Bibliotecas Digitais: uma proposta de aplicação**. Tese (Doutorado em Ciências da Informação) – Universidade Estadual Paulista, 2003.

SALES, R.; CAFÉ, L. Semelhanças e Diferenças entre Tesauros e Ontologias. **DataGramaZero**, Rio de Janeiro, v.9, n.4, ago. 2008.

SALTON, G.; YANG, C.S. On the especification of term values in automatic indexing. **Journal of the Americam Society for Information Science**, v.26, n.1, 1973.

SALTON, G.; McGILL, J.M. **Introduction to Modern Information Retrieval**. New York, McGraw-Hill, 1983.

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.

SARAWAGI, S. Information Extraction. **Foundations and Trends in Databases** v.1, n.3, 2008.

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

SOERGEL, D. The rise of ontologies or the reinvention of classification. **Journal of the American Society for Information Science**. v. 50, n. 12, 1999.

VICKERY, B. C. Ontologies. **Journal of Information Science.** v.23, n.4, 1997.