OntoSmart: um modelo de recuperação de informação baseado em ontologia

Edberto Ferneda

Graduação em Processamento de Dados pela antiga Fundação Educacional de Bauru (1985). Mestre em Informática pela Universidade Federal da Paraíb . Doutor em Ciências da Comunicação (Ciência da Informação) pela Universidade de São Paulo.Pós-doutorado pela Universidade Federal da Paraíba.Atualmente é professor do Departamento de Ciência da Informação da Universidade Estadual Paulista 'Julio Mesquita Filho' (UNESP) - Campus de Marília.

Guilherme Ataíde Dias

Graduado em Ciência da Computação pela Universidade Federal da Paraíba UFPB Campus II .Bacharel em Direito pelo Centro Universitário de João Pessoa UNIPE. Mestre em Organization & Management pela Central Connecticut State University CCSU. Doutor em Ciência da Informação (Ciências da Comunicação) pela Universidade de São Paulo USP e Pós-Doutor pela UNESP. Atualmente é professor Associado I na Universidade Federal da Paraíba. Vice-Coordenador Nacional do Grupo de Trabalho Informação e Tecnologia da ANCIB e Bolsista de Produtividade em Pesquisa (PQ) do CNPq.

http://dx.doi.org/10.1590/1981-5344/2081

Um sistema de recuperação de informação é um elemento mediador entre um estoque de informação e seus usuários. Sua eficácia depende do controle adequado da linguagem de representação dos itens de informações e das buscas de seus usuários. Este trabalho apresenta um modelo de recuperação de informação baseada em ontologia que usa a estrutura formal do modelo espaço vetorial. O vetor que representam um documento é criado durante o processo de indexação automático no qual uma ontologia fornece novos termos para enriquecer semanticamente essa representação. O vetor de busca é criado a partir de um processo de expansão de consulta, na qual novos termos são adicionados na expressão de busca inicialmente formulada pelo usuário a partir de

inferências na ontologia. Usando o modelo proposto, foi desenvolvido um sistema denominado OntoSmart, cujos resultados preliminares apontam em uma melhoria significativa na precisão dos resultados de busca.

Palavras-chave: *Recuperação de Informação baseada em Ontologia; Modelo Espaço Vetorial; Indexação Automática; Expansão de consulta*

OntoSmart: an ontology-based information retrievel model

Information retrieval system is a mediator element between a stock of information and its users. Its effectiveness depends on representation language of information items and requests of its users. This work presents an ontology-based information retrieval model which uses the formal structure of Vector Space Model. The vector that represent a document is created during the automatic indexing process, in which an ontology provide new terms in order to semantically enrich that representation. The search vector is created from a query expansion process, in which new terms are added in the search expression initially formulated by the user from inferences in the ontology. Using the proposed model, the OntoSmart system is being developed. Preliminary results show a significant improvement in the precision of search results.

Keywords: *Ontology-based Information Retrieval; Vector Space Model; Automatic Indexing; Query Expansion.*

Recebido em 23.04.2015 Aceito em 02.05.2017

1 Introdução

A eficiência de um sistema de recuperação de informação depende, por um lado, da linguagem de representação dos itens de informação (documentos) e, por outro lado, da forma como os usuários traduzem linguisticamente as suas necessidades de informação. Um sistema de recuperação de informação pode ser visto, portanto, como um elemento mediador entre um estoque de informação e os seus potenciais requisitantes (MEADOW *et al*, 2007, p.3). É tarefa do sistema de informação buscar formas de compatibilizar a terminologia utilizada na representação dos documentos e nas requisições dos usuários. Na Ciência

da Informação, as linguagens documentárias são tradicionalmente consideradas a ponte entre a informação e o usuário que a necessita. Para Fujita (2004), as linguagens documentárias são um conjunto controlado de termos que visam à representação de conceitos significativos de assuntos dos documentos utilizados na fase de indexação e busca. Elas proporcionam uma convergência entre a linguagem do indexador e a linguagem do usuário de um sistema de informação,

Para Tálamo, Lara e Kobashi (1992, p.197), as Linguagens Documentárias são tradicionalmente consideradas instrumentos de controle terminológico que atuam na representação da informação obtida pela análise e síntese de textos e na formulação das expressões de busca.

A ideia de agregar um controle terminológico a um sistema de recuperação de informação não é recente. Já na década de 1970, o professor e pesquisador Gerard Salton propunha métodos de construção de tesauros para serem utilizados em tais sistemas (SALTON, 1972). Na década de 1980, Salton e McGill propuseram a utilização de um tesauro no sistema SMART com o objetivo de incorporar novos termos de indexação aos termos previamente extraídos dos documentos por processos automáticos. Além disso, um tesauro poderia também ajudar o usuário a elaborar suas buscas. (SALTON; McGILL, 1983, p.75).

Na década de 1990 o termo "ontologia" começa a ser referenciado na área da Ciência da Computação e tomou notoriedade com o surgimento do projeto da Web Semântica. Muitos trabalhos tratam das diferenças e semelhanças entre tesauros e ontologias (JIMÉNEZ, 2004; SALES; CAFÉ, 2008; SALES; CAFÉ, 2009; CODINA; PEDRAZA-JIMÉNEZ, 2011; KLESS; MILTON, 2011). Como semelhanças, pode-se destacar que ambos têm como objetivo representar e compartilhar o vocabulário de um domínio do conhecimento a fim de possibilitar uma comunicação eficiente; as suas estruturas básicas são hierárquicas, agrupando termos ou conceitos em classes e subclasses e ambas podem ser utilizadas para organizar recursos informacionais. Segundo Qin e Paling (2000-01), a principal diferença entre tesauro e ontologia está no maior nível semântico das relações hierárquicas do tipo classe/subclasse e das relações "cruzadas" de uma ontologia.

A recuperação de informação baseada em ontologia (*ontology-based information retrieval*) já é um campo de pesquisa consolidado na Ciência da Computação, com um grande número dissertações e teses defendidas em diversos países. Tais trabalhos abordam uma diversificada gama de propostas e abordagens para a utilização de ontologias no processo de recuperação de informação.

Este trabalho propõe um modelo de recuperação de informação, denominado OntoSmart, no qual as ontologias, vistas como vocabulários controlados, são utilizadas como ferramentas de padronização do vocabulário tanto das representações dos documentos como das buscas dos usuários. Utiliza como alicerce teórico e prático o Modelo Espaço Vetorial, que permite incorporar diversos métodos e técnicas desenvolvidas ao longo de mais de quatro décadas de pesquisas nesse modelo.

2 Trabalhos relacionados

Muitas pesquisas sobre "recuperação de informação baseada em ontologia" estão em curso ou já apresentam resultados. Esses sistemas apresentam muitas características comuns, mas também podem diferir significativamente na maneira como as ontologias são utilizadas.

O sistema OntoSeek (GUARINO; MASOLO; VETERE, 1999) é um sistema de recuperação de informação baseado na descrição de produtos disponíveis em páginas amarelas e catálogo *on-line*. A descrição dos produtos e as consultas dos usuários são representados por meio de grafos conceituais derivados de ontologias. Assim o problema da recuperação de informação se reduz à equiparação (*mathcing*) de grafos. Os nós e arcos de um grafo que representa uma consulta são comparados aos nós e arcos de um grafo que representa um produto.

O sistema CIRI (AIRIO *et al*, 2004) utiliza ontologias na indexação de documentos e na criação e expansão de consultas. Inicialmente o usuário escolhe a ontologia relacionada ao seu interesse de busca e seleciona os termos em uma representação hierárquica e visual dos conceitos da ontologia escolhida. A partir de um conjunto inicial de termos escolhidos pelo usuário, o sistema expande automaticamente a consulta, considerando os relacionamentos entre os conceitos da ontologia.

O sistema OnAIR (PAZ-TRILLO; WASSERMANN; BRAGA, 2005) é um sistema de recuperação de trechos de vídeos a partir de consultas em linguagem natural. Foi testado utilizando um conjunto de entrevistas com a artista brasileira Ana Teixeira. Para esse objetivo foi desenvolvida uma ontologia sobre arte contemporânea. Os trechos de vídeo são indexados por meio de palavras-chave atribuídas por um especialista do domínio e por palavras contidas na transcrição do vídeo. A partir das consultas em texto livre, o sistema extrai termos relevantes e elimina palavras de pouca importância semântica. Para cada termo é atribuído um peso em função da frequência no *corpus* e de sua ocorrência na ontologia. A expansão das consultas é feita com a utilização dos conceitos e das relações da ontologia.

O sistema OWLIR (FININ *et al*, 2005) recupera documentos textuais contendo marcações semânticas provenientes do próprio texto e de uma ontologia. Tais marcações auxiliam no processo de indexação dos documentos, melhorando o desempenho da recuperação de informação. O sistema utiliza uma ontologia sobre eventos de uma universidade e foi aplicado sobre um *corpus* de páginas de anúncios de eventos desta mesma universidade. Inicialmente são extraídos termos das páginas visando identificar os tipos de eventos tratados na coleção. O sistema então anota as páginas utilizando informações extraídas dos textos acrescidas do conhecimento inferido na ontologia. Em seguida é realizada a indexação dos documentos anotados. A ontologia é utilizada também na expansão das consultas dos usuários.

O modelo que será apresentado neste trabalho possui muitas características semelhantes aos sistemas citados, porém se distingue por

uma abordagem relativamente mais simples e intuitiva na utilização de ontologias no processo de recuperação de informação.

3 Ontologia

A bibliografia da Ciência da Computação aponta que a primeira menção do termo ontologia em um trabalho da área se deu no artigo intitulado *"Another look at data"*, de George H. Mealy (1967). Desde então as ontologias têm despertado interesse de inúmeros pesquisadores da área. Porém, segundo Guizzardi (2005, p.56), somente a partir de 2001 é que se observa uma grande quantidade de trabalhos relacionados ao tema.

Gruber (1993) define ontologia com como uma "especificação formal explícita de uma conceitualização compartilhada". Por *formal* entende-se que esta especificação seja expressa num formato legível por computadores; *explícita* significa que os conceitos, as propriedades, as relações, as funções, as restrições e os axiomas devem estar formalmente definidos e passíveis de serem manipulados por computadores. Entende-se por *conceitualização* que tal representação seja referente a algum modelo abstrato de algum fenômeno do mundo real. Por *compartilhada*, compreende-se que esse conhecimento seja consensual (GRUBER, 1995; FENSEL, 2001; BORST, 1997).

Uma ontologia pode ser vista como um vocabulário de representação, geralmente especializado em algum domínio ou assunto, qualificado por conceituações de tipos de objetos e suas relações no mundo. É um corpo de conhecimento que descreve algum domínio, utilizando um vocabulário de representação (CHANDRASEKARAN; JOSEPHSON; BENJAMIN, 1999).

Segundo Jacob:

Ontologias são categorias de coisas que existem ou podem existir em um determinado domínio particular, produzindo um catalogo onde existem as relações entre os tipos e até os subtipos do domínio, provendo um entendimento comum e compartilhado do conhecimento de um domínio que pode ser comunicado entre pessoas e programas de aplicação. (JACOB, 2003, p.19).

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

Resumidamente, os componentes de uma ontologia são (RAMALHO, 2010, p.38):

> **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 a união de peças que formam uma estrutura na qual classes e subclasses definem um "esqueleto" hierárquico complementado por propriedades descritivas, propriedades relacionais, regras e axiomas. Toda classe é caracterizada por seus atributos; uma subclasse herda as características (atributos) da classe pai. Uma instância é a materialização de uma classe e representa um conceito ou uma entidade do mundo real. Quando uma classe é instanciada, cada um dos seus atributos pode então receber valores que irão individualizar aquele conceito ou entidade. É possível estabelecer regras que impõem restrições e limites às classes e atributos, e que se refletem nas suas instâncias. Uma ontologia é, enfim, uma estrutura conceitual que visa representar formalmente os conceitos e suas relações, regras e restrições lógicas de um determinado domínio.

Na Ciência da Informação, segundo Soergel (1999) e Vickery (1997), o termo ontologia começou a ser utilizado no final da década de 1990, principalmente por pesquisadores da área de Organização do Conhecimento. 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.

Ramalho (2010, p.38) acrescenta que:

Assim, as ontologias proporcionam liberdade para representar tipos de relacionamentos que não seriam possíveis em outros modelos de representação, podendo ser concebidas a partir de diversas técnicas de organização do conhecimento, cabendo aos desenvolvedores importantes decisões no momento da modelagem.

Enfim, as ontologias podem ser vistas como um novo instrumento a ser incorporado ao arsenal teórico e prático da Ciência da Informação. A aprendizagem de novos conceitos e novos recursos oferecidos pelas ontologias é um desafio para os profissionais da informação, mas que pode ser facilmente enfrentado utilizando toda bagagem teórica acumulada durante a história da Ciência da Informação.

4 Ontologia na recuperação de informação

De forma simplificada, o processo de recuperação de informação parte da comparação de dois elementos linguísticos: a representação dos documentos e a representação da expressão de busca. As ontologias se inserem no processo de recuperação de informação com o objetivo de prover um maior nível semântico de tais representações.

A representação dos documentos de um *corpus* é feita por meio da indexação, que visa descrever o conteúdo informacional de cada documento por meio de um conjunto de termos extraído do texto do próprio documento ou selecionado de um elemento auxiliar de padronização terminológica. As ontologias podem desempenhar um papel importante no processo de indexação por meio da disponibilização de uma estrutura conceitual e terminológica contextualizada em determinado domínio de conhecimento.

As necessidades de informação dos usuários é também um fator determinante para a eficiência de um sistema de recuperação de informação. A tradução da necessidade de informação em uma expressão de busca envolve elementos difíceis de serem formalizados. A utilização de uma ontologia no processo de especificação de buscas permite derivar novos termos e agregá-los automaticamente à expressão de busca inicialmente formulada pelo usuário, processo conhecido como "expansão de consulta".

Segundo Ferneda (2012, p.20), a eficiência de um sistema de recuperação de informação está ligada ao modelo que utiliza. Um modelo de recuperação de informação é a especificação formal de três elementos: a representação dos documentos, a representação da expressão de busca do usuário e a forma como esses dois elementos serão comparados por meio da função de busca. No modelo OntoSmart, utiliza-se o Modelo Espaço Vetorial como estrutura básica para a representação dos documentos e das expressões de busca, assim como diversas técnicas derivadas das pesquisas nesse modelo.

5 Modelo espaço vetorial

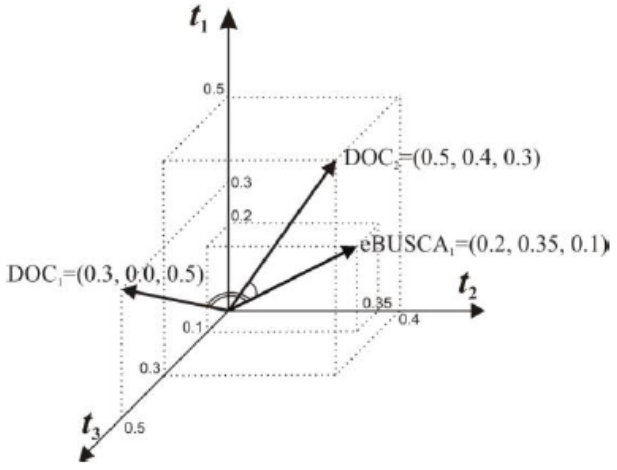
No Modelo Vetorial (Salton, 1968) um documento é representado por um vetor no qual cada elemento determina o peso ou a importância do respectivo termo na representação do conteúdo informacional do documento. Cada elemento do vetor é normalizado de forma a assumir valor entre zero (0) e um (1). Uma expressão de busca (consulta) é também representada por um vetor numérico onde cada elemento representa o grau de relevância do respectivo termo na necessidade de informação do usuário.

A Figura 1 apresenta um exemplo da representação vetorial de dois documentos (DOC1 e DOC_2) e uma expressão de busca ($eBUSCA_1$).

Figura 1 - Representação vetorial de dois documentos e uma expressão de busca

	t_1	t_2	t_3
$eBUSCA_1$	0.2	0.35	0.1

	t_1	t_2	t_3
DOC_1	0.3	0.0	0.5
DOC_2	0.5	0.4	0.3



Fonte: FERNEDA, 2012, p.34

A utilização de uma mesma representação tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade entre uma determinada busca e cada um dos documentos do *corpus*. Em um espaço vetorial contendo *N* dimensões, a similaridade (***sim***) entre um documento d_i e uma expressão de busca q é calculada utilizando a seguinte fórmula:

$$sim(d_i, q) = \frac{\vec{d}_i \bullet \vec{q}}{\left|\vec{d}_i\right| \times \left|\vec{q}\right|} = \frac{\sum_{k=1}^{N}(w_{k,i} \times w_{k,q})}{\sqrt{\sum_{k=1}^{N} w_{k,i}^2} \times \sqrt{\sum_{k=1}^{N} w_{k.q}^2}}$$

onde $w_{k,i}$ é o peso do *k*-ésimo elemento do vetor que representa o documento d_i e $w_{k,q}$ é o peso do *k*-ésimo elemento do vetor da expressão de busca q.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos do *corpus* são utilizados no ordenamento dos documentos resultantes. Esse ordenamento (*ranking*) permite agregar a um sistema alguns parâmetros que permitem restringir o resultado a um número máximo de documentos ou determinar um limite mínimo para o

valor da similaridade dos documentos resultantes de uma determinada busca.

6 Indexação automática baseada em antologia

O objetivo da indexação é sintetizar o conteúdo temático de um documento por meio de um conjunto de termos. Os termos de indexação são os "pontos de acesso" mediante os quais o documento é localizado e recuperado em um sistema de informação.

O processo de indexação realizada por seres humanos é dependente de critérios subjetivos e pessoais, relacionados à formação e à experiência do indexador. Assim, custo e a qualidade da indexação ficam fortemente atrelados a fatores não controláveis. Tais dificultadas levaram a estudos que buscavam soluções alternativas para auxiliar o indexador no exercício de sua atividade. As primeiras pesquisas em indexação automática aconteceram no final dos anos de 1950. O surgimento da Web fez deu um novo impulso nas pesquisas sobre indexação automática permanecesse praticamente constante.

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 representadas em linguagens legíveis por computador, o que permite a sua utilização nos mais variados processos computacionais.

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

7 Expansão de Consultas baseada em Ontologia

A eficiência de um sistema de recuperação de informação é dependente da especificação da busca do usuário. A variabilidade e imprecisão inerente ao ser-humano e às linguagens naturais, além do número reduzido de termos utilizados na especificação de sua expressão de busca, não permite uma interpretação exata e inequívoca de sua necessidade de informação.

A importância e as dificuldades do processo de especificação de buscas fez surgir na área de Recuperação de Informação um nicho de pesquisa em expansão de consulta (*query expansion*). Expansão de consulta é o termo utilizado para referenciar os métodos e processos que visam melhorar a eficiência da recuperação de informação baseados no pressuposto de que as consultas definidas pelos usuários muitas vezes não refletem suas reais necessidades de informação. O objetivo principal é

adicionar novos termos à consulta inicialmente formulada pelo usuário a fim de melhorar os resultados obtidos.

Uma ontologia pode ser utilizada na expansão de consultas por meio da inserção de novos termos derivados dos relacionamentos entre os seus conceitos. Além disso, a partir de uma interface adequada, as ontologias podem servir também como ferramentas para a seleção dos termos que irão compor a consulta do usuário. Isso permite que uma pessoa leiga em um determinado domínio ou assunto consiga realizar consultas pertinentes, ao mesmo tempo em que se familiariza com a terminologia do domínio de interesse.

8 O modelo OntoSmart

Utilizando-se conceitos de indexação baseada em ontologia e de expansão de consulta baseada em ontologia, foi desenvolvido um modelo e um sistema de recuperação denominado OntoSmart.

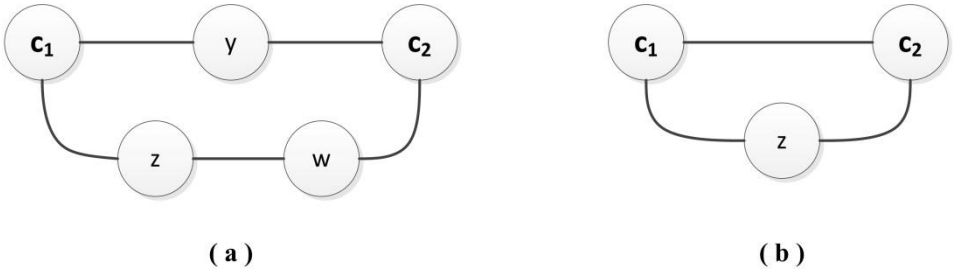
O modelo OntoSmart utiliza a estrutura formal do Modelo Espaço Vetorial, associado ao uso de ontologias como ferramenta de normalização da terminologia durante o processo de criação dos vetores representativos dos documentos e das buscas. A seguir serão definidos alguns conceitos básicos do modelo.

8.1 Distância semântica

Uma ontologia O = (C, R) é composta por um conjunto de conceitos $C=\{c_1, c_2,.., c_n\}$ interconectados por um conjunto de relacionamentos em $R=\{r_1, r_2, ..., r_n\}$. A distância semântica (**ds**) entre dois conceitos de uma ontologia (**c_1** e **c_2**) é igual ao número de relacionamentos existentes no menor caminho entre **c_1** e **c_2**.

Na Figura 2a, o menor caminho entre **c_1** e **c_2** é passando pelo conceito **y** e dois relacionamentos (**c_1, y)** e (**y, c_2)**. Portanto, *ds*(c_1, c_2)=2. Na Figura 2b *ds*(c_1, c_2)=1, pois os conceitos **c_1** e **c_2** são adjacentes, separados por um único relacionamento.

Figura 2 - Ilustração do conceito de distância semântica (ds)



Fonte: Elaborada pelos autores

O valor de ***ds*** entre um conceito e ele próprio é igual a zero. Assim, por exemplo, ***ds*(c1, c1)=0** e ***ds*(c2, c2)=0**.

8.2 Valor semântico

Tomando-se como referência um conceito ***c*** de uma ontologia, pode-se inferir que exista uma progressiva degradação do nível semântico dos conceitos a ele relacionados à medida que distância semântica (***ds***) vá aumentando.

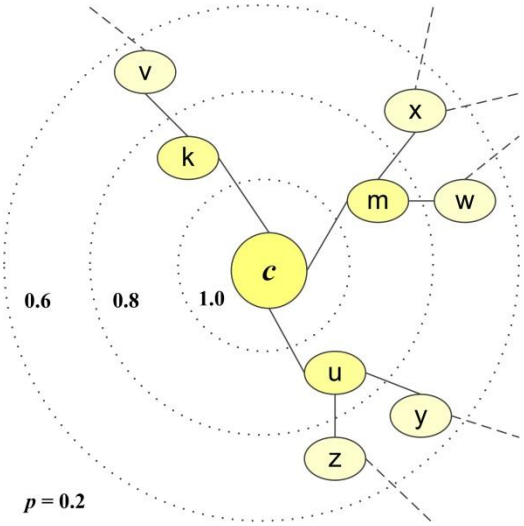
Dado um conceito **c** de uma ontologia, o valor semântico (**vs**) de cada um conceito ($\mathbf{c_i}$) é calculado da seguinte forma:

$$vs(c_i, c) = 1 - [\, ds(c_i, c) \times p \,]$$

onde **p** é um parâmetro numérico, entre 0 e 1, que define a diferença dos valores de **vs** a cada distância **ds** de um dado conceito $\mathbf{c_i}$ em relação a **c**.

Quanto maior a distância semântica (***ds***) do conceito central ***c***, menor será o valor semântico (***vs***) de um dado conceito da ontologia. O parâmetro ***p*** define a diferença dos valores de ***vs*** a cada valor de ***ds***. A Figura 3 apresenta uma ilustração mais ampla da aplicação dos conceitos de distância semântica (***ds***) e valor semântico (***vs***), considerando o parâmetro ***p***=0.2.

Figura 3 - Ilustração do conceito de valor semântico (*vs*)



Fonte: Elaborada pelos autores

A definição de um conceito central em uma ontologia faz surgir diversos níveis ou "camadas" concêntricas, onde cada camada é definida pela distância semântica (***ds***) em relação ao conceito central. Os conceitos de uma mesma camada recebem o mesmo valor semântico (***vs***). O conceito central possui ***vs*** igual a 1 e os demais conceitos terão ***vs*** menores, de acordo com a camada que ocupam e com o valor do parâmetro ***p***. Considerando **c** o conceito central e ***p***=0.2, os conceitos da Figura 3 terão os seguintes valores:

Conceito	*ds*	*vs*
c	0	1.0
k, m, u	1	0.8
v, x, w, y, z	2	0.6

Na Figura 3 foram apresentadas apenas três camadas de uma ontologia genérica. Com parâmetro ***p*** igual a 0.2, cada camada corresponde a um valor decrescente de ***vs,*** variando de 1 a 0.6. Considerando que ***vs*** não pode ser negativo e que uma ontologia pode ter um grande número de conceitos, é necessário definir um parâmetro ***k*** que limite o número de camadas a serem consideradas no cálculo de ***vs***.

Os valores dos parâmetros ***p*** e ***k*** são interdependentes. Não faz sentido, por exemplo, ***p*** = 0.2 e ***k*** = 8, pois isso acarretaria valores negativos de ***vs***. Portanto, o valor máximo que o parâmetro ***p*** pode assumir ($\boldsymbol{p}_{\max}$) é igual 1/***k***. De maneira formal temos:

$$p_{max} = \frac{1}{k}$$

No exemplo da Figura 3 o valor de ***k*** é igual a três (***k***=3). Portanto, o valor máximo que do parâmetro ***p*** pode assumir é 0.33 ($\boldsymbol{p}_{max}$=0.33).

8.3 Indexação dos documentos

No OntoSmart cada documento será representado por um único vetor numérico no qual cada elemento representa a importância (peso) do respectivo termo na representação do documento. Porém, diferentemente do Modelo Vetorial, no OntoSmart os pesos são calculados por meio da utilização da ontologia associada ao *corpus* do qual documento faz parte. A indexação de cada documento é realizada em duas fases: extração de termos e expansão dos índices.

Inicialmente será extrair do documento um conjunto de termos que irá representar o seu conteúdo informacional. Para cada termo é atribuído um valor numérico (peso) que expressa a relevância do respectivo termo na representação documento. A extração de termos e o cálculo de seus pesos são realizados por meio de um método de indexação por extração automática. Ao final desse processo poderia se obter, por exemplo, os seguintes termos de indexação e seus respectivos pesos:

Leão	**0.9**
Girafa	**0.85**
Zebra	**0.8**
Macaco	0.5
Floresta	0.4
Savana	0.35

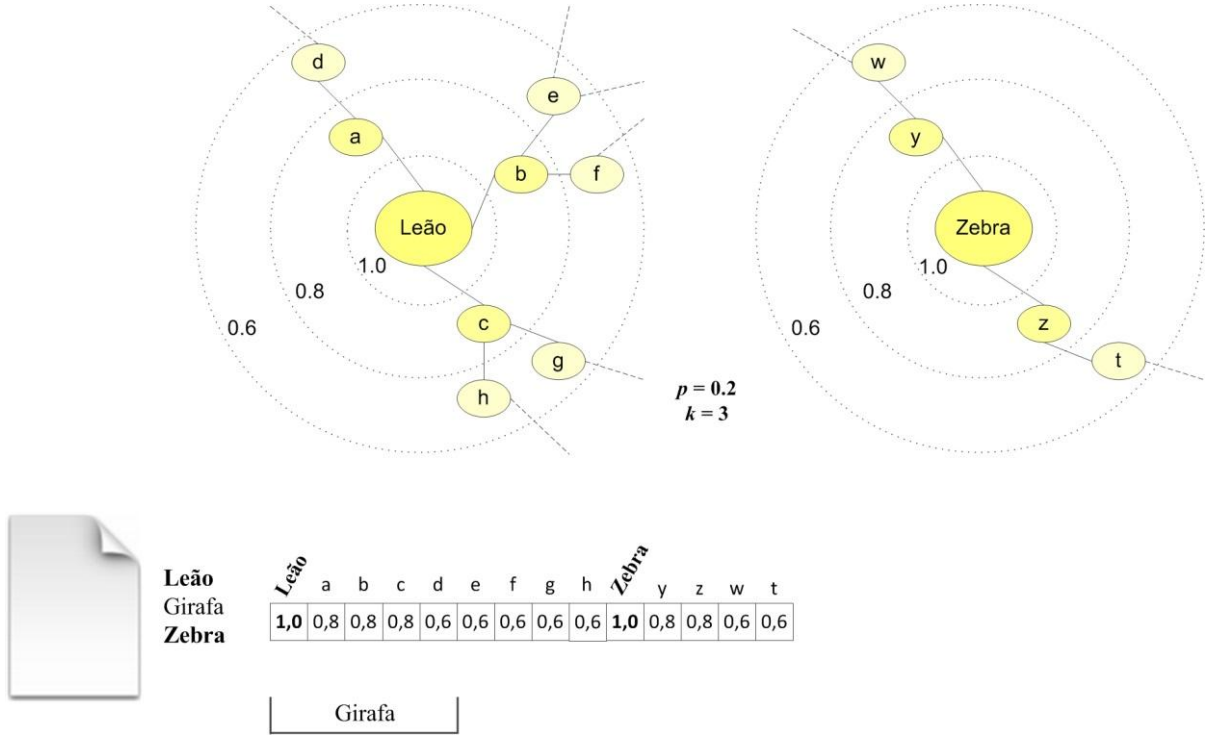
Nesse exemplo foram extraídos seis termos e considerados apenas aquele com peso igual ou superior a 0.8

No exemplo da Figura 4, verifica-se que apenas os termos "Leão" e "Zebra" têm relação com conceitos da ontologia. Assim, esses termos farão parte do vetor que representa o documento, com valor semântico (***vs***) igual a 1 (um). Os demais termos que irão compor o vetor do

documento serão derivados desses dois termos coincidentes, por meio suas relações.

Dado um determinado conceito, quanto maior a sua distante semântica (***ds***) do "conceito central", menor será o seu valor semântico (***vs***).

Figura 4 - Representação vetorial de um documento utilizando ontologia



Fonte: Elaborada pelos autores

Tomando-se "Leão" como conceito central da ontologia, deriva-se os demais termos de indexação, observando o valor de ***vs*** para cada camada da ontologia. A diferença dos valores de ***vs*** em cada camada da ontologia é dado pelo parâmetro ***p***, que no exemplo possui valor igual a 0.2. Assim, os termos da primeira camada adjacente (**a**, **b** e **c)** receberão o valor 0.8 e os termos **d**, **e**, **f**, **g** e **h** (segunda camada) terão valor igual a 0.6.

Considerando agora "Zebra" como o conceito central da ontologia, os conceitos **y** e **z** serão considerados termos de indexação do documento, ambos com ***vs*** igual a 0.8. Os conceitos **w** e **t** terão ***vs*** igual a 0.6.

O termo "Girafa" foi descartado por não estar representado por um conceito da ontologia. Porém, há de se considerar que esse termo foi extraído do texto do documento por um método estatístico que lhe atribuiu um peso de valor relativamente alto. Os termos extraídos do documento que não consta na ontologia serão armazenados em um tipo de repositório, formando um conjunto de potenciais conceitos que podem vir a ser inseridos na ontologia.

8.4 Expressão de busca

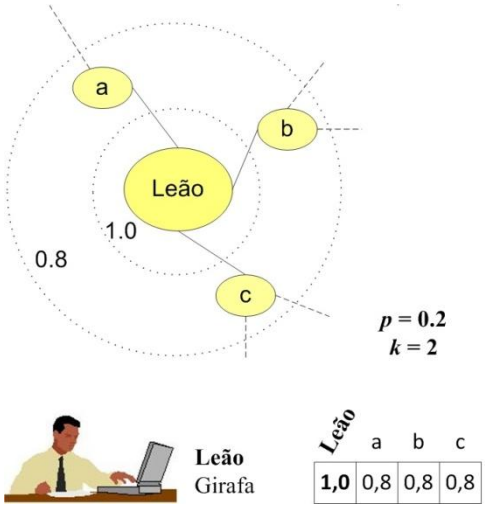
Uma expressão de busca é representada por um único vetor numérico no qual cada elemento corresponde à importância do respectivo termo para a descrição da necessidade de informação do usuário.

Antes da execução da busca, o usuário deve selecionar a ontologia do domínio ao qual se refere a sua necessidade de informação e definir valores para os parâmetros ***k*** e ***p*** para a expansão da sua busca/consulta, de forma similar à expansão dos termos de indexação.

Os termos definidos pelo usuário na sua expressão de busca (consulta) serão utilizados como conceitos centrais da ontologia associada a esse consulta. A ontologia terá duas funções: (1) expandir o conjunto de termos da consulta, acrescentando novos termos provenientes da ontologia; (2) atribuir pesos a cada um dos termos da consulta. Essas funções tomam como base a distância dos termos inicialmente definidos na busca/consulta e que se encontram diretamente representados na ontologia.

Considere uma consulta na qual o usuário utilizou dois termos: "Leão" e "Girafa". Fazendo-se uma busca na ontologia, verifica-se que apenas o primeiro termo está representado na ontologia. Assim, no vetor que representará esta consulta apenas o termo "Leão" estará presente com peso igual a 1 (um).

Figura 5 – Representação de uma expressão de busca



Fonte: Elaborada pelos autores

Tomando-se "Leão" como conceito central da ontologia e considerando os parâmetros ***p***=0.2 e ***k***=2, derivam-se os termos **a**, **b** e **c**, que farão parte da expressão de busca expandida. Tais termos receberão o valor 0.8, como exemplificado na Figura 5.

O termo "Girafa", que não está presente na ontologia, da mesma forma que nos documentos, será armazenado em um tipo de repositório. Se esse termo for utilizado muito frequentente, ele pode vir a ser integrado na ontologia, de acordo com critérios previamente determinados no sistema.

9 Considerações finais

No modelo de recuperação proposto, elementos linguísticos que formam uma ontologia são considerados termos de um vocabulário de domínio, utilizado como ferramenta de padronização terminológica das representações dos documentos e das buscas em um sistema de recuperação de informação. Tais representações utilizam como base formal o Modelo Espaço Vetorial, que fornece uma base matemática consistente e consolidada.

Uma vantagem evidente do modelo de recuperação proposto é a delimitação explícita do contexto no qual o processo de recuperação de informação é realizado. No OntoSmart todo documento faz parte de um *corpus* documental cujo domínio é definido pela ontologia a ele associada. Os documentos são indexados utilizando o vocabulário de domínio definido pelos conceitos dessa ontologia. Por sua vez, antes de expressar sua necessidade de informação o usuário define o seu domínio de interesse por meio da seleção de uma ontologia, que será utilizada para agregar novos termos à expressão de busca inicialmente formulada por ele. O Modelo Vetorial fornece a estrutura formal de representação tanto para os documentos como para as buscas, o que permite fornecer como resultado uma lista de documentos ordenados pelo grau de similaridade/relevância.

Percebe-se que o modelo OntoSmart possui uma simetria na forma de se construir os vetores que representam os documentos e os vetores das expressões de busca. Tal característica permitiu uma redução significativa do tempo e do esforço de desenvolvimento do sistema OntoSmart, pois permitiu racionalizar um considerável compartilhamento de código fonte.

O sistema OntoSmart, uma implementação do modelo de mesmo nome, ainda está em desenvolvimento mas já apresenta resultados bastante expressivos no que se refere à precisão dos resultado de uma busca. Alguns desses resultados estão apresentados em (FERNEDA; DIAS, 2013; NICOLINO; FERNEDA, 2014; PANSANI JUNIOR; FERNEDA, 2016).

Referencias

AIRIO, E. et al. CIRI: an ontology-based query interface for text retrieval. In: FINNISH ARTIFICIAL INTELLIGENCE CONFERENCE, 11., Finland, 2004. Web Intelligence. Proceedings… Vantaa, Finland, 2004.

BORST, W. N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. 227f. 1997. Tese (Doutorado) - Centre for Telematics for Information Technology, University of Twence, Enschede, 1997.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? IEEE Intelligent Systems, v. 14, n. 1, p. 20-26, 1999.

CODINA, L.; PEDRAZA-JIMÉNEZ, R. Tesauros y ontologías en sistemas de información documental. El profesional de la Información, v. 20, n. 5, p. 555-563, 2011.

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.

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, p. 93-107, 1996.

FENSEL, D. Ontologies: a silver bullet for knowledge management e electronic commerce. Berlin: Springer-Verlag, 2001.

FERNEDA, E.; DIAS, G.A. Um método de expansão automática de consulta baseada em ontologia. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 14., 2008, Florianópolis, SC. Anais… Florianópolis: ANCIB, 2008.

FERNEDA, E. Introdução aos Modelos Computacionais de Recuperação de Informação. Rio de Janeiro: Ciência Moderna, 2012.

FERNEDA, E.; DIAS, G. A. A lógica fuzzy aplicada à recuperação de informação. InterScientia, João Pessoa, v. 1, n. 1, p. 51-65, jan./abr. 2013.

FININ, T. et al. Information retrieval and the semantic web. In: ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES (HICSS'05), 38., 2005, Washington. Proceedings… Washington: IEEE Computer Society, 2005.

FUJITA M. S. L. A leitura documentária na perspectiva de suas variáveis: leitor-texto-contexto. DataGramaZero: Revista de Ciência da Informação, Rio de Janeiro, v. 5, n. 4, ago. 2004.

GARCÍA JIMÉNEZ, A. Instrumentos de representación del conocimiento: tesauros versus ontologías. Anales de Documentación, v. 7, p. 79-95, 2004.

GRUBER, T. A Translation approach to portable ontology specifications. Knowledge Acquisition, v. 6, n. 2, p. 199-220, 1993.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. International Journal Human-Computer Studies, v. 43, n. 5-6, p. 907-928, 1995.

GUARINO, N.; MASOLO, C.; VETERE, G. Ontoseek: content-based access to the web. IEEE Intelligent Systems, v. 14, n. 3, p. 70-80, 1999.

GUIZZARDI, G. Ontological foundations for structural conceptual models. The Netherlands: Universal Press, 2005.

JACOB, E. K. Ontologies and the semantic web. Bulletin of the American Society for Information Science and Technology, Silver Spring, p. 16-18, Apr./May 2003.

KLESS, D.; MILTON, S. Comparison of thesauri and ontologies from a semiotic perspective. In: AUSTRALASIAN ONTOLOGY WORKSHOP, 6., 2010, Australia. Conferences in research and practice in information technology: advances in ontologes. Adelaide; Australia: Australian Computer Society, 2010.

MEADOW, C. T. et al. Text Information Retrieval System. 3. ed. London UK: Elsevie, 2007.

MEALY, G. H. Another look at data. In: Proceedings. of the AFIPS'67 Fall Joint Computer Conference. Anaheim, CA. Washington, DC: Thomson Book, 1967.

NICOLINO, M. E. V. P. ; FERNEDA, E. Um método para a utilização de ontologias na indexação automática. Informação & Tecnologia (ITEC), v. 1, p. 13-33, 2014.

PANSANI JUNIOR, E. A.; FERNEDA, E. Ontologias no processo de indexação automática de documentos textuais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 17., 2016, Salvador. Anais... Salvador: ANCIB, 2016.

PAZ-TRILLO, C.; WASSERMANN, R.; BRAGA, P. P. An information retrieval application using ontologies. Journal of the Brazilian Computer Society, v. 11, n. 2, p. 17-31, 2005.

QIN, J.; PALING, S. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, v. 6, n. 2, 2000/2001.

RAMALHO, R. A. S. Desenvolvimento e utilização de ontologias em bibliotecas digitais: uma proposta de aplicação. 145 f. 2010. Tese (Doutorado em Ciências da Informação) - Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Campus de Marília, 2010.

SALES, R.; CAFÉ, L. Semelhanças e diferenças entre tesauros e ontologias. DataGramaZero, Rio de Janeiro, v. 9, n. 4, ago. 2008.

SALES, R.; CAFE, L. Diferenças entre tesauros e ontologias. Perspectivas em Ciência da Informação, v. 14, n. 1, p. 99-116, 2009.

SALTON, G. Automatic information organization and retrieval. New York: McGraw-Hill Book Company, 1968.

SALTON, G. Experiments in automatic thesaurus construction for information retrieval. In: FREIMAN, C. V.; GRIFFITH, J. E.; ROSENFELD, J. L. (Eds.). Information processing 71: proceedings of IFIP Congress 71. North-Holland, 1972. v.1.

SALTON, G.; McGILL, J. M. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

SOERGEL, D. The rise of ontologies or the reinvention of classification. Journal of the American Society for Information Science, v. 50, n. 12, p. 1119-1120, 1999.

TÁLAMO, M. F. G. M.; LARA, M. L. G.; KOBASHI, N. Y. Contribuição da terminologia para a elaboração de tesauros. Ciência da Informação, v. 21, n. 3, p. 197-200, 1992.

VICKERY, B. C. Ontologies. Journal of Information Science, v. 23, n. 4, p. 277-286, 1997.