OntoSmart: proposta de um modelo de recuperação de informação baseado em ontologia

OntoSmart: proposal for an ontology-based information retrieval model

Edberto FERNEDA (1), Guilherme Ataíde DIAS (2)

(1) Universidade Estadua Paulista – UNESP-Marília. Departamento de Ciência da Informação.
Av. Hygino Muzzi Filho, 737 - 17.525-000 – Marília - SP, ferneda@marilia.unesp.br

(2) Universidade Federal da Paraíba – UFPB. Departamento de Ciência da Informação
Campus I - Jardim Cidade Universitária - 58.051-900 – João Pessoa - PB, guilhermeataide@gmail.com

Resumo

Um sistema de recuperação de informação é um ambiente linguístico mediador da comunicação entre um estoque de informação e seus requisitantes. Sua eficiência depende de um controle adequado da representação dos documentos e das requisições de seus usuários. No contexto da Ciência da Informação o tesauro se consolidou como uma ferramenta bastante eficiente na representação da informação. Porém, as ontologias surgem como uma nova tecnologia que auxiliam na representação e na organização da informação e do conhecimento. Esta pesquisa tem por objetivo desenvolver um modelo de recuperação de informação baseado em ontologia que utiliza como estrutura formal o Modelo Espaço Vetorial. Os vetores que representam os documentos são criados durante o processo de indexação automática. A partir de um conjunto inicial de termos extraídos dos documentos, procede-se uma inferência na ontologia com a finalidade de enriquecer a indexação. De forma semelhante, a expressão de busca de um usuário é também representada por um vetor, criado a partir de um processo de expansão de consulta por meio de inferências em uma ontologia. Utilizando o modelo proposto, deu-se início ao desenvolvimento de um sistema de recuperação de informação denominado OntoSmart, cujos resultados parciais apontam para um aumento significativo na precisão da recuperação.

Palavras-chave: Recuperação de Informação; Ontologia; Modelo Espaço Vetorial; Indexação Automática; Expansão de Consulta.

Abstract

An information retrieval system is a linguistic environment that mediates the communication between a stock of information and its requesters. Their efficiency depends on an adequate control of representation of the documents and of the requests of its users. In the context of Information Science, thesaurus has established as a very efficient tool in the representation of information. However, ontologies are emerging as a new technology that assist the representation and organization of information and knowledge. This research aims to develop an ontology-based information retrieval model that uses the formal structure of the Vector Space Model. The vectors representing the documents are created during the automatic indexing process. From an initial set of terms extracted from the documents there shall be an inference in the ontology in order to enrich the indexing. Similarly, a user's query is also represented by a vector created from a query expansion process through inferences in ontology. Using the proposed model, it was initiated the development of an information retrieval system called OntoSmart, whose partial results point to a significant increase in accuracy of retrieval.

Keywords: Information retrieval; Ontology; Vector Space Model; Automatic Indexing; Query Expansion.

1. Introdução

Um sistema de recuperação de informação é um ambiente linguístico cuja eficiência depende de um controle adequado da representação dos itens de informação e das requisições de seus usuários. Insere-se como um elemento mediador da comunicação entre um estoque de informação e os seus potenciais requisitantes.

Em seu papel de mediador de um processo comunicativo, é tarefa do sistema definir um código ou uma linguagem comum entre emissor e receptor, entre os conteúdos informacionais dos documentos e as requisições dos usuários. Na Ciência da Informação, tradicionalmente as linguagens documentárias são utilizadas como um elemento comutador entre a informação e o usuário que a necessita. Para Fujita (2004), as linguagens documentárias são um conjunto

Ferneda, Edberto; Dias, Guilherme Ataíde. OntoSmart: proposta de um modelo de recuperação de informação baseado em ontologia. En XII Congreso ISKO España y II Congreso ISKO España-Portugal, 19-20 de noviembre, 2015, Organización del conocimiento para sistemas de información abiertos. Murcia: Universidad de Murcia.

controlado de termos que visam representar os assuntos tratados pelos documentos e utilizados na fase de indexação e busca. Essas linguagens promovem a convergência entre a linguagem do indexador e a linguagem do usuário de um sistema de recuperação informação.

Tálamo, Lara e Kobashi (1992, p.197) apontam que:

> As Linguagens Documentárias são tradicionalmente consideradas instrumentos de controle terminológico que atuam em dois níveis: a) na representação da informação obtida pela análise e síntese de textos; b) na formulação de equações de busca da informação.

Na década de 1970, Salton (1972) já propunha métodos de construção de tesauros para serem utilizados em sistemas de recuperação de informação. Segundo Salton e McGill (1983, p.75), apresentado por meio de uma interface adequada, um tesauro poderia ajudar o usuário a elaborar suas buscas, ao mesmo tempo em que o familiariza com o vocabulário utilizado pelo sistema.

A partir da década de 1990 o termo "ontologia" começa a ser frequentemente referenciado na área da Ciência da Computação. O tema tomou notoriedade ainda maior e se expandiu para outras áreas com o surgimento do projeto da Web Semântica, na qual as ontologias aparecem como parte de destaque na sua estrutura.

Muitos trabalhos tratam das diferenças e semelhanças entre tesauros e ontologias (Codina; Pedraza-Jiménez, 2011; Kless; Milton, 2011; Sales; Café, 2009; Jiménez, 2004;). Dentre as semelhanças, pode-se destacar que: (1) ambos têm como objetivo representar e compartilhar os conceitos ou o vocabulário de um domínio a fim de possibilitar uma comunicação eficiente; (2) as suas estruturas básicas são hierárquicas, agrupando termos ou conceitos em categorias e subcategorias (classes e subclasses); (3) ambas podem ser utilizadas para catalogar ou organizar recursos informacionais.

Na Ciência da Computação, a recuperação de informação baseada em ontologia (*ontology-based information retrieval*) já é um campo de pesquisa consolidado, 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.

O sistema CIRI (AIRIO *et al*, 2004) utiliza ontologias na indexação de documentos, 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 OntoSeek (Guarino; Masolo; Vetere, 1999) é um sistema de recuperação de informação baseado na descrição de produtos disponíveis em páginas amarelas e catálogos *on-line*. A descrição dos produtos e as consultas dos usuários são representadas por meio de grafos conceituais derivados de ontologias. Assim, o problema de recuperação de informação se reduz à equiparação (*matching*) 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 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.

Esse 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 sistema FROM (Pereira; Ricarte; Gomide, 2006) implementa o modelo ontológico relacional *fuzzy* para recuperação de informação textual. O sistema faz a expansão da consulta considerando as relações existentes em uma ontologia de domínio composta por categorias e palavras-chaves. As categorias denotam os conceitos mais gerais e as palavras-chaves denotam conceitos mais específicos. Uma consulta do usuário pode ser composta apenas por palavras-chaves, por categorias ou por ambas. A expansão da consulta é feita pela adição de novas categorias e palavras-chaves, em função das conexões existentes na ontologia. A similaridade dos documentos em relação à consulta é calculada por meio de operações *fuzzy*, e são recuperados os documentos que apresentarem similaridade acima de um determinado valor.

Este trabalho propõe um modelo de recuperação de informação baseado em ontologia, 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 (Salton, 1968), que permite incorporar diversos métodos e técnicas desenvolvidas ao longo de mais de quatro décadas de pesquisas nesse modelo. O OntoSmart possui muitas características semelhantes aos sistemas citados, porém se distingue por uma abordagem relativamente simples e intuitiva.

2. Ontologia

Até a última década do século XX, ontologia era considerada primariamente uma disciplina da Filosofia. Atualmente, as ontologias têm sido utilizadas de diferentes maneiras em diversas áreas, e vêm ocupando cada vez mais a atenção de estudiosos da Ciência da Informação e da Ciência da Computação, tendo em vista a possibilidade de melhorar significativamente a representação de um domínio do conhecimento.

O tema tem despertado o interesse de inúmeros pesquisadores da área, principalmente após a criação de fóruns temáticos tal como a série de conferências FOIS (*Formal Ontology and Information Systems*), em meados da década de 1990. Porém, somente a partir de 2001 é que se observa uma grande quantidade de trabalhos relacionados ao tema (Guizzardi, 2005, p.56).

Uma definição de ontologia muito citada é a de Gruber (1993), que descreve uma ontologia 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 considerada um vocabulário de representação, geralmente especializado em algum domínio ou assunto, qualificado por conceituações de tipos de objetos e suas relações no mundo. Em outras palavras, é um corpo de conhecimento que descreve algum domínio, utilizando um vocabulário de representação (Chandrasekaran; Josephson; Benjamin, 1999).

Segundo Jacob (2003, p.19):

> 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.

Uschold (1996) ressalta a necessidade de explicitação dos relacionamentos entre os termos de uma ontologia:

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

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 inanima-

dos 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 na forma 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.

Toda classe é caracterizada por seus atributos ou propriedades. 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. 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.

O autor acrescenta ainda que:

> 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 (Ramalho, 2010, p.37).

Outro fator determinante para a distinção das ontologias e os modelos de representação tradi-

cionalmente utilizados na Ciência da Informação é a própria natureza dos relacionamentos. Para possibilitar a realização de inferências automáticas é necessário que as relações expressas na ontologia sejam formalmente descritas, rotuladas e categorizadas. 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.

As ontologias se colocam 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.

3. Ontologias na Recuperação de Informação

Existem dois elementos linguísticos que afetam diretamente na eficiência de um sistema de recuperação de informação: 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 das representações dos documentos e das necessidades de informação dos usuários.

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.

A representação adequada das 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. Um usuário não familiarizado com a terminologia de uma área do conhecimento ou de um determinado assunto de seu interesse tenderá a expressar sua necessidade de informação utilizando termos muito genéricos ou coloquiais, o que pode resultar na recuperação de um número excessivo de documentos não relevantes. A utilização de uma ontologia no processo de especificação de buscas permite derivar novos termos e agregá-los automaticamente à expressão de busca inicial do usuário, em um processo conhecido como “expansão de consulta” (*query expansion*).

Segundo Ferneda (2012, p.20):

> Um modelo de recuperação de informação é a especificação formal de três elementos: a representação dos documentos, a representação da expressão de busca e como esses dois elementos serão comparados, a função de busca. A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que utiliza, influenciando diretamente em seu modo de operação.

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.

4. Modelo Espaço Vetorial

Como mencionado anteriormente, o modelo OntoSmart tem como base o 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 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.

5. Indexação Automática baseada em Ontologia

A indexação de um documento visa representar o seu conteúdo informacional por meio de um conjunto de termos com o objetivo de sintetizar o seu conteúdo, ressaltando o que lhe é essencial. Um termo de indexação é constituído de uma ou mais palavras cujo significado remete pretensamente a um conceito único, não ambíguo. Os termos de indexação servem também como pontos de acesso mediante os quais o documento é localizado e recuperado em um sistema de informação.

Embora a prática da indexação possa ser regulada por políticas e princípios institucionais, o processo de indexação realizada por seres humanos (manual) é dependente de critérios subjetivos e pessoais relacionados à formação e à experiência do indexador. Assim, o tempo despendido e a qualidade da indexação ficam fortemente atrelados a fatores não controláveis, o que pode afetar o custo desse processo.

As dificuldades inerentes à indexação manual e a grande quantidade de documentos derivados da explosão informacional após a Segunda Gerra, justificaram 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, época de rápido desenvolvimento da Ciência da Computação. O surgimento dos microcomputadores na década de 1980 e da Web nos anos de 1990 fez com que o nível de interesse nas pesquisas sobre indexação automática permanecesse praticamente constante até os dias de hoje.

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 e a eficácia do processo de recuperação de informação.

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 representadas em linguagens processáveis 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.

6. Expansão de Consultas baseada em Ontologia

Embora importante para uma recuperação eficiente, a especificação da busca (consulta) é dependente do usuário, com toda a variabilidade e imprecisão inerente ao ser-humano. Além disso, geralmente as buscas dos usuários são expressas por meio de um número reduzido de termos, não permitindo uma interpretação exata e inequívoca da necessidade de informação do usuário.

A importância e as dificuldades do processo de especificação de buscas fizeram surgir na área de Recuperação de Informação um campo de pesquisa em expansão de consulta. Expansão de consulta (*query expansion*) 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 de busca 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 assunto consiga realizar consultas pertinentes, ao mesmo tempo em que se familiariza com a terminologia do domínio de interesse.

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.

7. O modelo 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.

7.1. 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**.

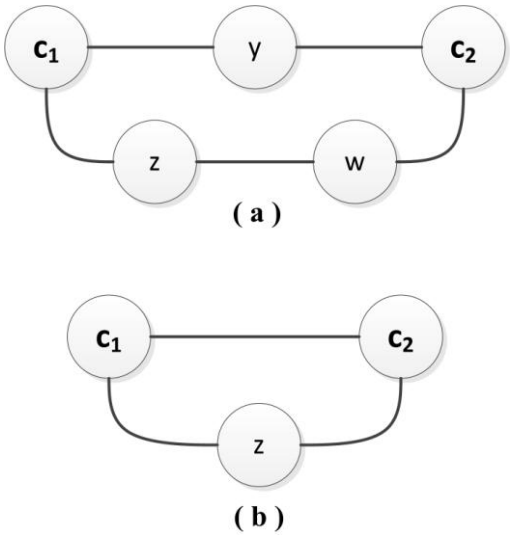


Figura 1. *Ilustração do conceito de distância semântica (ds)*

Na Figura 1a, 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 1b *ds*(c_1, c_2)=1, pois os conceitos **c_1** e **c_2** são adjacentes, separados por um único relacionamento.

O valor de ***ds*** entre um conceito e ele próprio é igual a zero. Assim, por exemplo, *ds*(c_1, c_1)=0 e *ds*(c_2, c_2)=0.

7.2. 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 a distância semântica (***ds***) vá aumentando.

Dado um conceito **c** de uma ontologia, o valor semântico (***vs***) de cada conceito (***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 ***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 2 apresenta uma ilustração da aplicação dos conceitos de distância semântica (***ds***) e valor semântico (***vs***), considerando o parâmetro ***p***=0.2.

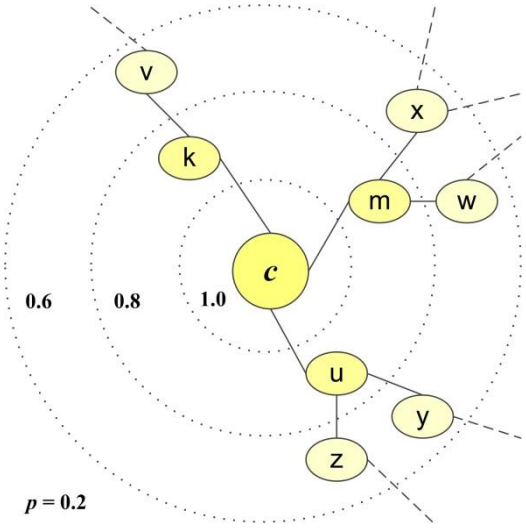


Figura 2. *Ilustração do conceito de valor semântico (vs)*

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 2 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 2 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 o valor de ***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 (p_{max}) é igual 1/***k***. De maneira formal temos:

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

No exemplo da Figura 2 o valor de ***k*** é igual a três (***k***=3). Portanto, o valor máximo que o parâmetro ***p*** pode assumir é 0.33 (p_{max}=0.33).

7.3. Indexação dos documentos

No OntoSmart um documento é 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 modelo OntoSmart os pesos são calculados por meio da ontologia associada ao documento. A indexação de cada documento é realizada em duas fases: extração de termos e expansão dos índices.

Inicialmente será extraído do documento um conjunto de termos que represente o seu conteúdo informacional. Para cada termo é atribuído um valor numérico (peso) que expressa a relevância do respectivo termo na representação do conteúdo informacional do documento. A extração de termos e o cálculo de seus pesos são realizados por meio de algum método 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 do documento seis termos, mas considerados apenas aqueles com peso igual ou superior à um determinado valor, no caso 0.8. Dessa forma, o documento será representado por três termos: “Leão”, “Girafa” e “Zebra”, que serão utilizados para realizar inferências em uma ontologia.

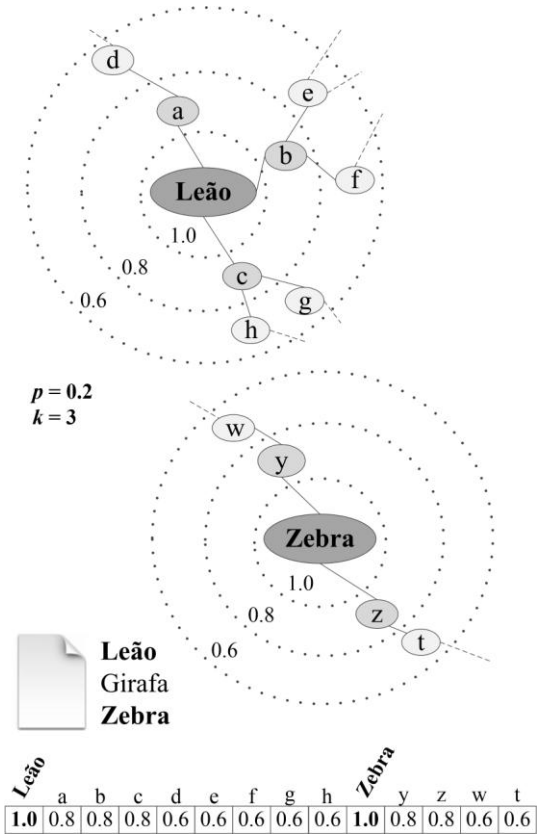


Figura 3 - Representação vetorial de um documento utilizando ontologia

Na Figura 3, pode-se verificar 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.

Tomando-se “Leão” como conceito central, 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 é dado pelo parâmetro ***p***, que no exemplo possui valor igual a 0.2. Assim, os termos **a**, **b** e **c** receberão o valor 0.8. Os termos **d**, **e**, **f**, **g** e **h** terão valor igual a 0.6.

Considerando agora “Zebra” como o conceito central, 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 método estatístico que lhe atribuiu um peso relativamente alto. Esses termos serão armazenados em um tipo de repositório, formando um conjunto de potenciais conceitos a serem inseridos na ontologia.

7.4. Expressão de busca

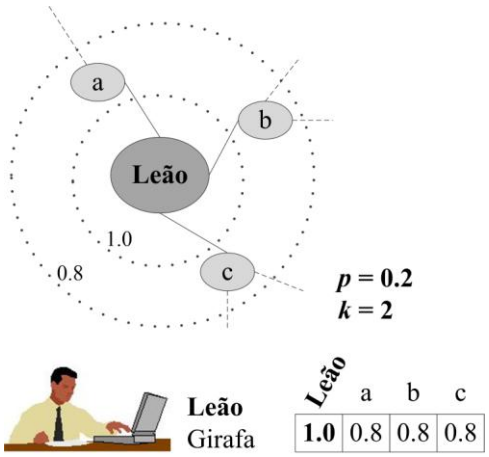
Uma expressão de busca é representada por um único vetor numérico no qual cada elemento corresponde à importância do respectivo termo para a descrição da necessidade de informação do usuário.

Antes da execução da busca, o usuário deve selecionar a ontologia do domínio ao qual se refere a sua necessidade de informação. Os parâmetros ***k*** e ***p*** devem estar previamente definidos antes da expansão da sua busca.

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

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

Tomando-se “Leão” como conceito central 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 4.



Leão	a	b	c
1.0	0.8	0.8	0.8

Figura 4. Representação de uma expressão de busca

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 frequentemente utilizado pelos usuários ele pode vir a ser integrado na ontologia, de acordo com critérios previamente determinados em um sistema.

8. 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 o domínio de um documento é explicitamente definido por meio de sua relação com uma determinada ontologia. 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 resultados de uma busca. Alguns desses resultados foram apresentados por Ferneda e Dias (2013).

9. Agradecimentos

Este trabalho é parte do resultado de pesquisa de pós-doutorado financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por meio do projeto PROCAD-NF-099/2009.

Referências

AIRIO, E.; JÄRVELIN, K.; SAATSI, P.; KEKÄLÄINEN , J.; SUOMELA, S. CIRI – an ontology-based query interface for text retrieval. In: HYVÖNEN, E.; KAUPPINEN T.; SALMINEN, M.; VILJANEN, L.; ALA-SIURU, P. (Eds) Proceedings of the 11th Finnish Artificial Intelligence Conference, 2004.

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

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

CODINA, L.; PEDRAZA-JIMÉNEZ, R. Tesauros y Ontologías en Sistemas de Información Documental. El profesional de la Información, v.20, n.5, 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.

MEADOW, C.T.; BOYCE, B.R.; KRAFT, D.H.; BARRY, C. (2007). Text Information Retrieval System. 3rded. London UK: Elsevie, 2007.

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

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.

DING, Y.; FOO, S. Ontology research and development. Part 1- a review of ontology generation. Journal of Information Science, v.28, n. 2, 2002.

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.

FININ, T.; MAYFIELD, J.; JOSHI, A.; COST, R.S.; FINK, C. Information retrieval and the semantic web. In: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05). Washington: IEEE Computer Society, 2005.

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

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

FERNEDA, E.; DIAS, G.A. Um Método de Expansão Automática de Consulta Baseada em Ontologia. In: XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). Florianópolis-SC, 2013.

FUJITA M. S. L. (2004). 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.

GRUBER, T. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, v.6, n.2, 1993.

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

GUARINO, N.; MASOLO, C.; VETERE, G. Ontoseek: Content-based access to the web. IEEE Intelligent Systems, v.14, n.3, 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, apr./may,, 2003.

JIMÉNEZ, A.G. Instrumentos de Representación del Conocimiento: tesauros versus ontologias. Anales de Documentación, n.7. Universidad de Murcia, 2004.

KLESS, D.; MILTON, S. Comparison of thesauri and ontologies from a semiotic perspective. In: Proceedings of the Sixth Australasian Ontology Workshop. Conferences in Research and Practice in Information Technology. Advances in Ontologes. Adelaide, Australia: Australian Computer Society, 2010.

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

PEREIRA, R.; RICARTE, I.; GOMIDE, F. Fuzzy relational ontological model in information search systems. In: SANCHEZ, Elie (Ed.). Fuzzy Logic and The Semantic Web, p.395–412, Elsevier B.V.: Amsterdan, 2006.

QIN, J.; PALING, S. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, v.6, n2, 2000-01.

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, 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, 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, v.1. North-Holland, 1972.

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, 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, 1992.

USCHOLD, M. Building Ontologies: Towards a Unified Methodology. proceedings of Expert Systems. In: Annual Conference of the British Computer Society Specialist Group on Expert Systems, 16, p.16-18 December, 1996, Cambridge. *Anais...,*Cambridge, 1996

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