UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB
Centro de Ciências Sociais Aplicadas – CCSA
Departamento de Ciência da Informação – DCI

Edberto Ferneda

Ontologia como recurso de padronização terminológica em um Sistema de Recuperação de Informação

João Pessoa
Março/2013

Edberto Ferneda

Ontologia como recurso de padronização terminológica em um Sistema de Recuperação de Informação

Relatório de Pesquisa apresentado ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba - UFPB, em cumprimento às exigências do estágio pós-doutoral.

Pós-doutorando: Edberto Ferneda

Supervisor: Guilherme Ataíde Dias

João Pessoa
Março/2013

Agradecimentos

A elaboração do projeto PROCAD-NF-099/2009 e a realização deste trabalho não teria sido possível sem a colaboração de muitas pessoas, mas de forma especial agradeço aos seguintes pesquisadores:

PPGCI/UNESP

Profa. Dra. Plácida Leopoldina Ventura Amorim da Costa Santos

Profa. Dra. Silvana Aparecida Borsetti Gregório Vidotti

Profa. Dra. Maria José Jorente

PPGCI/UFPB

Prof. Dr. Guilherme Ataíde Dias

Profa. Alba Ligia de Almeida Silva

Profa. Dra. Bernardina Juvenal Freire

Profa. Dra. Isa Maria Freire

Prof. Dr. Gustavo Henrique de Araújo Freire

Tendo a lua aquela gravidade aonde o homem flutua
Merecia a visita não de militares,
Mas de bailarinos

Herbert Vianna

Resumo

Desde a década de 1950, a importância dos sistemas de recuperação de informação cresce em função da quantidade de informação disponível. Apesar do acelerado avanço tecnológico observado nesse período, a busca por informações relevantes e úteis é ainda uma tarefa árdua. Recuperar informação envolve, por um lado, um acervo documental que deve ser representado por expressões linguísticas que resumem seu conteúdo informacional. Por outro lado, temos seres humanos que tentam descrever linguisticamente as suas necessidades de informação a fim de obterem documentos relevantes para satisfazer tais necessidades. Portanto, um sistema de recuperação de informação é um ambiente linguístico mediador na comunicação entre um estoque de informação e seus requisitantes. Sua eficiência depende de um controle adequado da linguagem de representação dos itens de informação e das requisições de seus usuários. Este trabalho apresenta um modelo de recuperação de informação baseado em ontologia que utiliza como estrutura formal o Modelo Espaço Vetorial. Os vetores que representam os documentos e as buscas são criados a partir de uma ontologia de domínio, utilizada como elemento de normalização terminológica. Os vetores dos documentos são criados durante o processo de indexação automática, no qual as ontologias fornecem novos termos a fim de enriquecer semanticamente a representação dos documentos. O vetor de busca é criado a partir de um processo de expansão de consulta, no qual novos termos são inseridos na expressão de busca inicialmente formulada pelo usuário a partir de inferências realizadas em uma ontologia. Utilizando o modelo proposto, está sendo desenvolvido o sistema OntoSmart. Após a conclusão de um primeiro protótipo totalmente funcional, será possível a realização de testes comparativos para verificar a sua eficiência e eficácia. Contudo, o OntoSmart será utilizado como base para futuras pesquisas em recuperação de informação.

Abstract

Since the 1950s, the importance of information retrieval systems increases as a function of the amount of information available. Despite the rapid technological advancement observed during this period, the search for relevant and useful information is still an arduous task. Retrieve information involves a collection of documents that must be represented by linguistic expressions that summarize their informational content, and in the other side users trying to describe linguistically their information needs in order to obtain relevant documents to meet such needs. Therefore, a information retrieval system is a linguistic environment mediating the communication between a stock of information and its users. Its effectiveness depends on adequate control of language for representation 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 vectors representing documents and queries are created from a domain ontology, which is used as an element of terminology standardization. The documents vectors are created during the automatic indexing process, in which the ontologies provide new terms in order to semantically enrich those representations. 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. Upon completion of the first fully functional prototype, it will be possible to carry out comparative tests to verify its efficiency and effectiveness. However, OntoSmart will be used as the basis for future research in information retrieval.

Lista de Figuras

Lista de Quadros

Sumário

1
Introdução

Os estoques de informação prontamente disponíveis e compartilhados por meios eletrônicos aumentam continuamente, fazendo também aumentar a importância dos sistemas de recuperação de informação. Desde as primeiras pesquisas, na década de 1950, até o surgimento da Web, no início dos anos de 1990, o papel de tais sistemas passou de simples ferramentas experimentais a sistemas de uso geral, úteis a todos que necessitem de informação para suas atividades. Em mais de meio século de pesquisas, aliado a um acelerado avanço das tecnologias de informação e comunicação, inúmeras ideias, conceitos e técnicas de recuperação de informação foram propostos e desenvolvidos. Porém, a busca por informações relevantes e úteis é ainda uma tarefa bastante árdua. Essa situação leva a refletir sobre os reais elementos envolvidos no processo de recuperação de informação, que aparentemente são alheios aos avanços tecnológicos, ou pelos menos às tecnologias atualmente disponíveis.

A recuperação de informação envolve, por um lado, um acervo documental composto de itens informacionais. Por outro lado, temos pessoas, seres humanos com as mais variadas necessidades de informação que buscam por documentos relevantes para satisfazer tais necessidades. Recuperar informação implica, portanto, em operar seletivamente um estoque de informação, o que envolve processos cognitivos difíceis de serem formalizados. A utilização de recursos computacionais nessa tarefa parte de inevitáveis simplificações teóricas e de adequações de conceitos subjetivos tais como “relevância“ e “necessidade de informação”, além do próprio conceito de informação.

Um sistema de recuperação de informação é um ambiente linguístico cuja eficiência depende de um controle adequado da linguagem de representação dos itens de informação e

das requisições de seus usuários. Insere-se como um agente mediador na comunicação entre um estoque de informação e os seus potenciais requisitantes. Nesse sentido, Meadow *et al* afirmam que:

> Recuperação de informação é um processo de comunicação. Em certo sentido é um meio pelo qual autores e criadores de registros se comunicam com os leitores, mas indiretamente e possivelmente com um longo intervalo de tempo entre a criação de uma mensagem ou texto e a sua entrega para o usuário de um sistema de recuperação de informação. Às vezes o sistema de recuperação de informação ou bibliotecário conduzindo uma pesquisa pode passar informações sobre o provável relevância ou valor do que é recuperado, aumentando a informação para o conjunto de itens recuperados. Os registros de uma base de dados são criados e montados sem conhecimento exatamente de quem irá lê-los, ou sob quais circunstâncias. As linguagens e os canais de tal sistema de comunicação são bastante diferentes de outros modelos bem conhecidos, tais como a radiodifusão ou a comunicação ponto-a-ponto (MEADOW *et al*, 2007, p.3, tradução nossa).[1]

No seu papel de mediador de um processo comunicativo, é tarefa do sistema definir um código, uma linguagem comum entre emissor e receptor, entre os conteúdos informacionais dos documentos e a requisições dos usuários. Na Ciência da Informação, as linguagens documentárias são tradicionalmente consideradas como a ponte entre a informação e o usuário que a necessita. Cintra (2002) afirma que a construção dessas linguagens visa às atividades de indexação, armazenamento e recuperação da informação. Fujita (2004) aponta que 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. Proporcionam uma convergência entre a linguagem do indexador e a linguagem do usuário de um sistema de informação, "já que vários autores podem utilizar diferentes palavras para expressar uma mesma ideia, assim como os usuários podem apresentar diversidade de vocabulário quando da expressão de uma estratégia de busca". Campos (2001) apresenta uma ideia mais genérica a respeito das Linguagens Documentárias, definindo-as como instrumentos utilizados para representar o conhecimento de uma determinada área do saber.

[1] IR is a communication process. In one sense it is a means by which authors or creators of records communicate with readers, but indirectly and with a possibly long time lag between creation of a message or text and its delivery to the IRS user. Sometimes the IRS or librarian conducting a search may pass on information about the probable relevance or value of what is retrieved, thereby adding information to the set of retrieved items. The records of a database are created and assembled without knowledge of exactly who will read them, or under what circumstances. The languages and channels of such a communication system are quite different from other well-known models, such as broadcasting or point-to-point communication.

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.

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 puramente matemáticos. Apresentado por meio de uma *interface* adequada, um tesauro pode também ajudar o usuário a elaborar suas buscas, ao mesmo tempo em que o familiariza com o vocabulário utilizado pelo sistema. (SALTON; McGILL, 1983, p.75).

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 (camada) de destaque na sua estrutura.

Ainda recentemente muitos trabalhos tratam das diferenças e semelhanças entre tesauros e ontologias (CODINA; PEDRAZA-JIMÉNEZ, 2011; KLESS; MILTON, 2011; SALES; CAFÉ, 2008; 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. No entanto, segundo Qin e Paling (2000-01), as ontologias se caracterizam por um maior nível semântico das relações hierárquicas do tipo classe/subclasse e das relações "cruzadas". Ding e Foo (2002) destacam que uma ontologia permite a comunicação entre humanos e computadores enquanto que os vocabulários controlados, criados no contexto da biblioteconomia, são ferramentas utilizadas para facilitar a comunicação entre seres humanos.

Moreira (2003, p.97) aponta para as origens e propósitos distintos para os dois instrumentos. O tesauro "nasceu como um instrumento prático para auxiliar na indexação e

busca de documentos", uma aplicação mais direcionada aos especialistas. As ontologias nasceram da "necessidade de descrever os objetos digitais e suas relações", uma aplicação mais direcionada aos procedimentos automatizados, às inferências computacionais por meio de agentes inteligentes. O ponto comum entre os dois instrumentos refere-se ao fato de estarem relacionados com a descrição ou a representação de alguma coisa.

Em um sistema de recuperação de informação, o nível e a precisão das representações dos itens de informação e das necessidades de informação dos usuários afetam diretamente no desempenho do sistema. Embora originalmente o propósito das ontologias se distancie dos objetivos de uma linguagem documentária, o seu poder de representação as tornaram uma opção natural para solução de alguns problemas relacionados à recuperação de informação.

Observando a produção bibliográfica ligada à Ciência da Computação percebe-se que 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.

Este trabalho propõe um modelo de recuperação de informação no qual as ontologias 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, o que permite incorporar diversos métodos e técnicas desenvolvidas ao longo de mais de três décadas de pesquisas nesse modelo.

1.1 Hipótese de Pesquisa

Neste trabalho as ontologias são vistas como ferramentas de padronização terminológica das representações dos documentos e das buscas dos usuários de um sistema de recuperação de informação. Essa padronização tem por objetivo a obtenção de melhores resultados no processo de recuperação de informação. Neste contexto, "melhores resultados" compreende principalmente a obtenção de documentos relevantes, que efetivamente atendam às necessidades de informação dos usuários. Isto é, espera-se uma melhoria significativa dos índices de revocação e precisão dos resultados das buscas.

A hipótese que se enuncia é:

As ontologias, vistas como uma forma de linguagem documentária, possibilitam a melhoria dos resultados obtidos na recuperação de informação por meio da compatibilização da terminologia utilizada na representação dos documentos (índices) e na representação das necessidades de informação dos usuários (expressões de busca)

1.2 Objetivos

1.2.1 Geral

Propor um modelo de recuperação de informação baseado no Modelo Espaço Vetorial, que utiliza ontologias de domínio como um elemento normalizador e unificador da linguagem de representação dos documentos e das buscas dos usuários.

1.2.2 Específicos

- Realizar um levantamento histórico do termo "ontologia" e estudar as formas de apropriação do conceito de ontologia pela Filosofia, Ciência da Computação e Ciência da Informação;
- Apresentar a área de Recuperação de Informação (*Information Retrieval*), seus principais modelos e as formas de inserção das ontologias no processo de recuperação de informação;
- Fazer um levantamento bibliográfico do tema "recuperação de informação baseado em ontologia" e dos subtemas: "indexação automática baseada em ontologia" e "expansão de consulta baseada em ontologia".
- Fazer um estudo sobre os métodos de indexação automática e as formas de utilização de ontologias como elemento normalizados dos termos de indexação;
- Realizar um levantamento dos métodos de expansão de consulta, com ênfase à utilização de ontologia nesse campo de pesquisa.
- Desenvolver um sistema computacional de recuperação de informação que utilize ontologias como elemento normalizador das representações tanto dos documentos como das necessidades de informação dos usuários.

1.3 Delimitação do tema de pesquisa

A utilização de elementos de normalização terminológica em sistemas de recuperação de informação remonta à década de 1960. Diversas técnicas e métodos utilizam listas de palavras, dicionários, tesauros ou outra forma de léxico. Há apenas pouco mais de uma década o conceito de ontologia se tornou tema comum nos meios acadêmicos, principalmente nas áreas da Ciência da Computação e Ciência da Informação.

Esse trabalho ficará restrito ao estudo da utilização de ontologias em sistemas de recuperação de informação. Serão, portanto, referenciados preferencialmente trabalhos que abordam esse tema e trabalhos que descrevem sistemas que utilizam especificamente ontologias, desconsiderando aqueles que utilizam tesauros ou outra estrutura conceitual.

1.4 Organização do trabalho

O **Capítulo 2** apresenta um histórico do termo "ontologia", seu significado e seus diferentes usos na Filosofia, na Ciência da Computação e na Ciência da Informação.

No **Capítulo 3** define "Recuperação de Informação" como uma área de pesquisa de interesse comum entre a Ciência da Informação e Ciência da Computação. São apresentados os elementos envolvidos no processo de recuperação de informação e os chamados modelos "clássicos". Por fim, serão apresentadas as principais formas de utilização de ontologias no processo de recuperação de informação.

Por um lado, a utilização de uma ontologia no processo de recuperação de informação se efetiva por meio da agregação termos/conceitos às representações dos documentos de um *corpus* durante o processo de indexação. O **Capítulo 4** define o processo de indexação e indexação automática, assim como as principais formas de automação do processo de indexação utilizando ontologias.

Por outro lado, a eficiência de um sistema de recuperação de informação é dependente da terminologia utilizada pelos usuários nas representações de suas necessidades de informação (expressões de busca). As ontologias assumem um papel importante na padronização terminológica das buscas, na medida em que fornecem um vocabulário específico de um determinado domínio, reduzindo ambiguidades semânticas. Os termos relacionados aos conceitos de uma ontologia podem ser agregados às expressões de busca dos

usuários durante o processo de "expansão de consulta". No **Capítulo 5** são detalhados os principais métodos de expansão de consulta e o processo de expansão de consulta utilizando uma ontologia.

No **Capítulo 6** é apresentado o sistema OntoSmart, desenvolvido no transcorrer desse trabalho. Esse sistema utiliza ontologia tanto no processo de indexação como no processo de expansão de consultas a fim de compatibilizar as representações dos documentos (índices) com a representação da necessidade de informação dos usuários (expressão de busca).

Por fim, no **Capítulo 7** serão apresentadas algumas observações e considerações finais sobre o presente trabalho.

1.5 Da terminologia utilizada

O tema principal deste trabalho, Recuperação de Informação, envolve pelo menos dois campos científicos: a Ciência da Informação e a Ciência da Computação. Surge daí problemas terminológicos decorrentes das diferentes nomenclaturas utilizadas para referenciar um mesmo conceito.

Considerando que este trabalho parte de interesses de investigação em Ciência da Informação, a terminologia utilizada será preferencialmente dessa área. Porém, alguns termos provenientes da Ciência da Computação já estão consolidados e são amplamente utilizados em diversos domínios científicos. Nesse caso a preferência recai sobre o termo mais comumente utilizado.

1.6 Trabalhos relacionados

Embora seja um tema recente, muitas pesquisas sobre *recuperação de informação baseada em ontologia* estão em curso ou já apresentam resultados substanciados em diversos sistemas. Esses sistemas apresentam muitas características comuns, mas também podem diferir significativamente na maneira como as ontologias são utilizadas.

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 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 de recuperação de informação se reduz à equiparação (*mathcing*) de grafos. Os nós e arcos de um grafo que representa uma consulta, sã comparados aos nós e arcos de um grafo 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.

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 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.

O sistema OntoSmart (Capítulo 6) 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. O sistema OntoSmart tem como "alicerce" o Modelo Espaço Vetorial, no qual o nível de representatividade/relevância (peso) dos termos de indexação e dos termos de buscas são valorados. As ontologias são vistas como vocabulários controlados que possibilitam uma unificação ou padronização da terminologia utilizada nas representações dos documentos e das buscas dos usuários por meio da agregação novos termos e seus respectivos pesos derivados de inferências sobre os conceitos de uma ontologia.

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.

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

CAMPOS, M. L.A. **Linguagem documentária***:* teorias que fundamentam sua elaboração. Niterói: EDUFF, 2001.

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

CODINA, L.; PEDRAZA-JIMÉNEZ, R. Tesauros y Ontologías en Sistemas de Información Documental. **El profesional de la Información**, v.20, n.5, 2011.

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

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)**. 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.

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.

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

MOREIRA, Alexandra. **Tesauros e ontologias**: estudo de definições presentes na literatura das áreas das ciências da computação e da informação, utilizando-se o método

analíticosintético. 2003. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte, 2003.

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

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.

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

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

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.

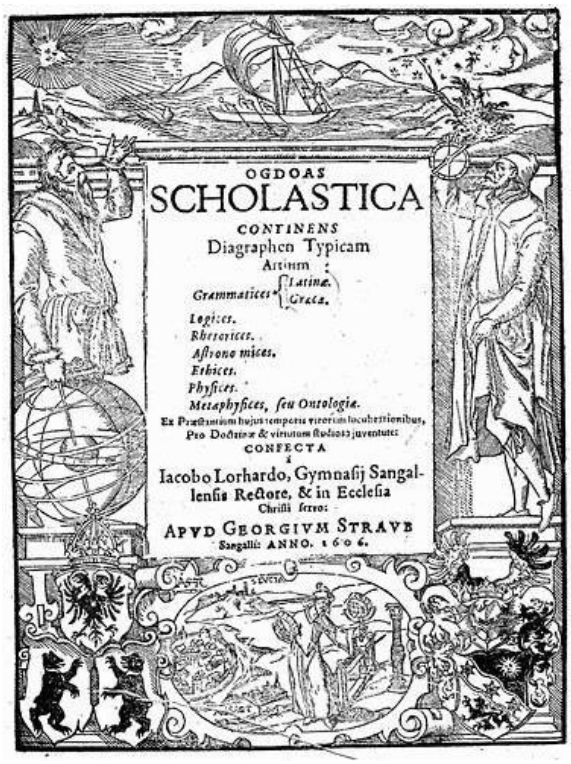
SUOMELA, S.; KEKÄLÄINEN, J. User evaluation of ontology as query construction tool. **Information Retrieval**, v.9, n.4, 2006.

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.

2
Ontologia

Neste capítulo será apresentado um breve histórico do surgimento do termo ontologia, seu significado e seus diferentes usos, primeiramente na sua área de origem, a Filosofia, e em seguida na Ciência da Computação e na Ciência da Informação.

Figura 2.1 – Capa do livro *Ogdoas Scholastica*, de 1606



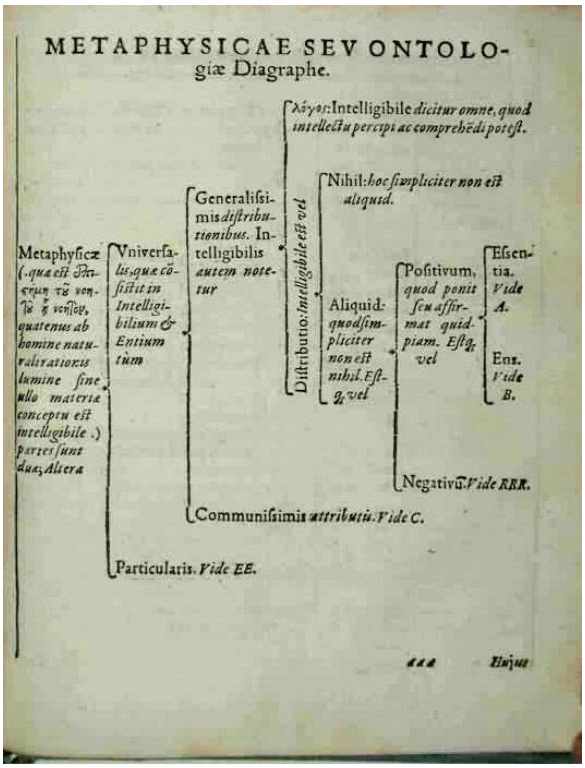
Fonte: http://readtiger.com/wkp/en/Jacob_Lorhard, Acessado em 16.09.2012

A primeira menção do termo ontologia é atribuída ao filósofo e pedagogo Jacob Lorhard (*Jacobus Lorhardus*) (1561-1609) em sua obra *Ogdoas Scholastica*, de 1606. Como pode ser observado na Figura 2.1, *Ogdoas* era um volume dividido em 8 livros sobre

Gramática do Latim, Gramática do Grego, Lógica, Retórica, Astronomia, Ética, Física e Metafísica (ou ontologia), respectivamente.

O título do livro 8, "*Metaphysices seu ontologiæ*", indica que os termos ontologia e metafísica são utilizados como sinônimos. Lorhard mostra sua antologia de forma diagramática (Figura 2.2), assemelhando-se a um hipertexto (ØHRSTRØM; SCHÄRFE; UCKELMAN, 2008).

Figura 2.2 – Página do livro *Ogdoas Scholastica*

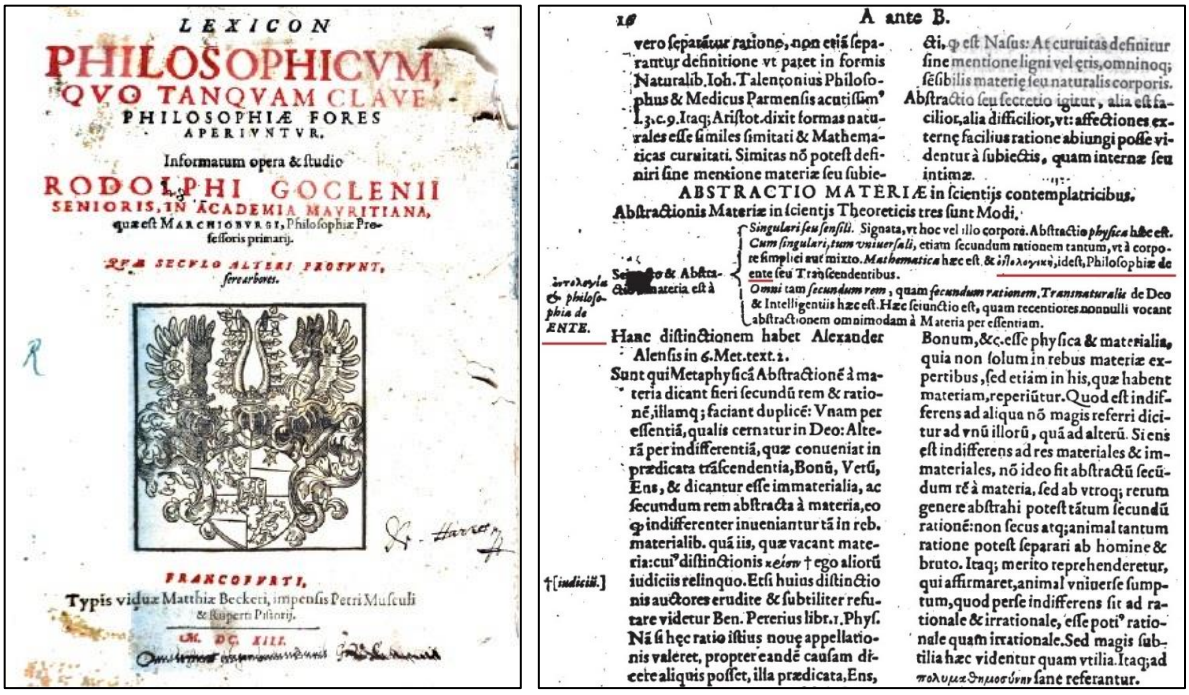


Fonte: Theory and History of Ontology, by Raul Corazzon
(http://www.ontology.co/jacob-lorhard.htm). **Acessado em 25/10/2012**

Um ano após a publicação de *Ogdoas Scholastica*, Jacob Lorhard, morador na cidade de St. Gallen (Suiça), recebeu uma oferta para se tornar professor de teologia em Marburg (Alemanha). É possível que nesse período Lorhard tenha conhecido Rudolph Göckel (*Goclenius*) (1547-1628) que era professor de lógica, ética e matemática em Marburg. Uma hipótese plausível é que Lorhard e Göckel tenham se encontrado algumas vezes durante o ano de 1607, compartilhando suas descobertas. Por alguma razão, a estada de Lorhard em Marburg foi breve, retornando à St. Gallen em menos de um ano. Jabob Lorhard faleceu em 19 de maio de 1609 com aproximadamente 47 anos (ØHRSTRØM; SCHÄRFE; UCKELMAN, 2008).

Em 1613 foi publicada uma segunda edição revisada do livro de Lorhard sob o título *Theatrum Philosophicum*. Nesta edição, a palavra "ontologia" não aparece na capa, mas permaneceu no interior do volume. Nesse mesmo ano foi publicado o livro de Rudolph Göckel intitulado *Lexicon Philosophicum* (Figura 2.3), onde na margem esquerda da página 16 aparece a grafia grega para a palavra ontologia – οντολογία – seguida de sua definição: "*philosophia de ENTE"*. No corpo do texto pode ser lido: οντολογία*, idest, Philosophiæ de ente seu de Transcendentibus* (LIMA-MARQUES, 2006, p.33).

Figura 2.3 – Capa e página 16 do livro *Lexicon Philosophicum*, de 1613



LEXICON
PHILOSOPHICVM,
QVO TANQVAM CLAVE
PHILOSOPHIÆ FORES
APERIVNTVR.
Informatum opera & ſtudio
RODOLPHI GOCLENII
SENIORIS, IN ACADEMIA MAVRITIANA,
quæ eſt MARCHIORVM, Philoſophiæ Profeſſoris primarij.
RES A SECVLO ALTERI PROSVNT,
ſeruabones.
FRANCOFVRTI,
Typis viduæ Matthiæ Beckeri, impenſis Petri Muſculi
& Ruperti Prioris.
CIↃ. DC. XIII.

16 A ante B.

vero ſeparatur ratione, non etiã ſeparantur definitione vt patet in formis Naturalib. Ioh. Talentonius Philoſophus & Medicus Parmenſis acutiſſim⁹ 1.3.c.9. Itaq; Ariſtot. dixit formas naturales eſſe ſimiles ſimitati & Mathematicas curuitati. Simitas nõ poteſt definiri ſine mentione materiæ ſeu ſubie-

cti, q eſt Naſus: At curuitas definitur ſine mentione ligni vel æris, omninoq; ſẽſibilis materiæ ſeu naturalis corporis. Abſtractio ſeu ſecretio igitur, alia eſt facilior, alia difficilior, vt: affectiones externę facilius ratione abiungi poſſe videntur à ſubiectis, quam internæ ſeu intimæ.

ABSTRACTIO MATERIÆ in ſcientijs contemplatricibus.

Abſtractionis Materiæ in ſcientijs Theoreticis tres ſunt Modi.

Singulari ſeu ſenſili. Signata, vt hoc vel illo corpore. Abſtractio *phyſica* hãc eſt.
Cum ſingulari, tum vniuerſali, etiam ſecundum rationem tantum, vt à corpore ſimplici aut mixto. *Mathematica* hæc eſt. & *τὸ λογικὸν*, idest, Philoſophiæ de ente ſeu Transcendentibus.
Omni tam *ſecundum rem*, quam *ſecundum rationem*, *Transnaturalis* de Deo & Intelligentiis hæc eſt. Hæc ſeiunctio eſt, quam recentiores nonnulli vocant abſtractionem omnimodam à Materia per eſſentiam.

Seiunctio & Abſtractio materiæ eſt à

ὀντολογία & philoſophia de ENTE.

Hanc diſtinctionem habet Alexander Alenſis in 6. Met. text. 2.

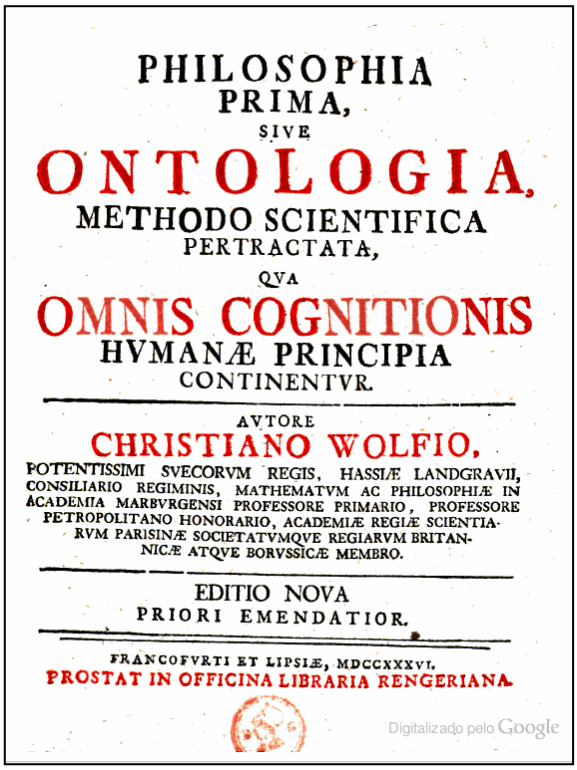
Sunt qui Metaphyſicã Abſtractionẽ à materia dicant fieri ſecundũ rem & rationẽ, illamq; faciant duplicẽ: Vnam per eſſentiã, qualis cernatur in Deo: Alterã per indifferentiã, quæ conueniat in prædicata trãſcendentia, Bonũ, Verũ, Ens, & dicantur eſſe immaterialia, ac ſecundum rem abſtracta à materia, eo q; indifferenter inueniantur tã in reb. materialib. quã iis, quæ vacant materia: cuius diſtinctionis κρίσιν † ego aliorũ iudiciis relinquo. Etſi huius diſtinctionis auctores erudite & ſubtiliter refutare videtur Ben. Pererius libr. 1. Phyſ. Nã ſi hęc ratio iſtius noue appellationis valeret, propter eandẽ cauſam dicere aliquis poſſet, illa prædicata, Ens, Bonum, &c. eſſe phyſica & materialia, quia non ſolum in rebus materiæ expertibus, ſed etiam in his, quæ habent materiam, reperiũtur. Quod eſt indifferens ad aliqua nõ magis referri dicitur ad vnũ illorũ, quã ad alterũ. Si ens eſt indifferens ad res materiales & immateriales, nõ ideo fit abſtractũ ſecũdum rẽ à materia, ſed ab vtroq; rerum genere abſtrahi poteſt tãtum ſecundũ rationẽ: non ſecus atq; animal tantum ratione poteſt ſeparari ab homine & bruto. Itaq; merito reprehenderetur, qui affirmaret, animal vniuerſe ſumptum, quod perſe indifferens ſit ad rationale & irrationale, eſſe poti⁹ rationale quam irrationale. Sed magis ſubtilia hæc videntur quam vtilia. Itaq; ad πολυμαθημοσύνην ſane referantur.

†[*indiciii.*]

Fonte: Google Books (http://books.google.com.br). Acessado em 03/12/2012

Foi apenas no ano de 1730, com a publicação da obra *Philosophia prima sive Ontologia* (Figura 2.4), de Christian Wolff (1679-1754), que o termo ontologia tomou visibilidade nos círculos filosóficos, sendo considerado sinônimo de *metaphysica generalis* – parte da metafísica que analisa as características do ser em geral. O livro de Wolff propõe investigar os predicados mais gerais de todos os entes por meio de um "método demonstrativo", racional e dedutivo.

Figura 2.4 – Capa do livro *Philosophia prima sive Ontologia*, de 1730


PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QVA
OMNIS COGNITIONIS
HVMANÆ PRINCIPIA
CONTINENTVR.
AVTORE
CHRISTIANO WOLFIO,
POTENTISSIMI SVECORVM REGIS, HASSIÆ LANDGRAVII, CONSILIARIO REGIMINIS, MATHEMATVM AC PHILOSOPHIÆ IN ACADEMIA MARBVRGENSI PROFESSORE PRIMARIO, PROFESSORE PETROPOLITANO HONORARIO, ACADEMIÆ REGIÆ SCIENTIARVM PARISINÆ SOCIETATVMQVE REGIARVM BRITANNICÆ ATQVE BORVSSICÆ MEMBRO.
EDITIO NOVA
PRIORI EMENDATIOR.
FRANCOFVRTI ET LIPSIÆ, MDCCXXXVI.
PROSTAT IN OFFICINA LIBRARIA RENGERIANA.
Digitalizado pelo Google

Fonte: Google Books (http://books.google.com.br). Acessado em 03/12/2012

Chauí (2012, p.229) apresenta de forma detalhada a etimologia e o significado da palavra ontologia:

> Essa palavra é composta de duas outras: *onto* e *logia*. *Onto* deriva de dois substantivos gregos, *tà onta* ("os bens e as coisas realmente possuídas por alguém"; e "as coisas realmente existentes"). *Tà onta* deriva do verbo *ser*, que, em grego, se diz *einai*. O particípio presente desse verbo se diz *on* (sendo, ente). Dessa maneira, as palavras *tà onta* ("as coisas") e *on* ("ente") levaram a um substantivo: *tò on*, que significa "o Ser". O Ser é o que é realmente e se opõe ao que *parece ser*, à aparência. Assim, *ontologia* significa "estudo ou conhecimento do Ser, dos entes ou das coisas tais como são em si mesmas, real e verdadeiramente, correspondendo ao que Aristóteles chamara de Filosofia Primeira, isto é, o estudo do Ser enquanto Ser"

Castro (2008, p.7) define a palavra ontologia de forma mais simplificada:

> Ela é o resultado da junção de dois termos gregos *onta* (entes) e *logos* (teoria, discurso, palavra). Ao pé da letra, ontologia significa, portanto, teoria dos entes. "Ente" está aí representando todas as coisas sobre as quais se pode dizer que são – ou que a ontologia é a teoria do ser enquanto tal.

Talvez pela dificuldade na tradução dos elementos constitutivos da palavra "ontologia", é possível encontrar na literatura algumas variações de sua etimologia. Dentre os

dicionários *on-line* da língua portuguesa, iDicionário Aulete[2] é o que traz a definição mais completa da palavra ontologia:

> (on.to.lo.gi.a)
> sf.
> 1. **Fil.** Parte da filosofia que estuda a natureza dos seres, o ser enquanto ser.
> 2. **Fil.** Doutrina sobre o ser.
> 3. **Hist. Med.** Doutrina segundo a qual os fenômenos patológicos têm existência própria, não tendo relação com fenômenos fisiológicos.
> 4. **Inf.** Campo da informática que trata de conceitualizar de forma explícita e formal (portanto processável por máquina e compartilhável) conceitos e restrições elacionados a certo domínio de interesses.
> [F.: *ont(o)*- + -*logia*.]

Considerada em qualquer de seus aspectos, uma ontologia possui a função de fornecer uma forma de organização dos seres e as coisas, o mundo, a realidade, o conhecimento.

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 Computação e da Ciência da Informação, tendo em vista a possibilidade de melhorar significativamente a representação de um domínio do conhecimento.

2.1 Ontologia na Filosofia

Na Filosofia, o termo ontologia possui sua origem na Metafísica, a "Filosofia Primeira" de Aristóteles (384-322 a.C.), que trata do estudo do ser em sua essência. Embora Aristóteles nunca tenha utilizado o termo ontologia, a ideia de uma ciência do ser remete à sua obra *Categorias*, texto que abre o *Oragon* – conjunto de textos relacionados à lógica. Apesar de composta de apenas um livro, costuma-se dividir o conteúdo desta obra em duas partes: a primeira parte, chamada de *Prædicamenta*, se estende do capítulo I ao IX e é considerada genuinamente aristotélica. A segunda parte, que se estende do capítulo X ao XV, chamada de *Post-Prædicamenta*, não há certeza se a autoria é de Aristóteles ou de seus discípulos.

[2] http://aulete.uol.com.br/ontologia. Acessado em 15/01/2013.

Segundo Chauí (2012, p.139), Aristóteles define os termos ou categorias como "aquilo que serve para designar uma coisa". As dez categorias Aristotélicas são apresentadas e exemplificadas no Quadro 2.1:

Quadro 2.1 – Categorias de Aristóteles

Categoria	Exemplos
Substância	Homem, Sócrates, animal
Quantidade	Dois metros de comprimento
Qualidade	Branco, grego, agradável
Relação	O dobro, a metade, maior do que
Lugar	Em casa, na rua, no alto
Tempo	Ontem, hoje, agora
Posição	Sentado, deitado, de pé
Posse	Armado (tendo armas)
Ação	Corta, fere, derrama
Paixão ou passividade	Está cortado, está ferido

Fonte: Chauí (2012, p.139)

Provavelmente devido à dificuldade de tradução do original grego para línguas latinas, é possível encontrar na literatura variações dos rótulos das categorias. A categoria "Posse", por exemplo, é assim denominada por Chauí e diversos outros autores. Porém, há autores que preferem o termo "Hábito"; outros utilizam a denominação "Estado". Ainda em relação à categoria "Posse", ela pode ser denominada também de "Ação permanente", sendo que nesse caso a categoria "Ação" é chamada de "Ação transeunte" (ARISTÓTELES, 2004, p.106). Enfim, nas pesquisas realizadas para este trabalho encontrou-se variações importantes não só quanto à denominação (rótulo) das categorias, mas também na ordem de apresentação e na interpretação de cada categoria.

Com as dez categorias aristotélicas busca-se classificar o que pode ser dito sobre qualquer coisa. Chauí (2012, p.139) explica que:

> As categorias ou termos indicam o que uma coisa é ou faz, ou como está. São aquilo que nossa percepção e nosso pensamento captam imediata e diretamente numa coisa, não precisando de qualquer demonstração, pois nos dão a apreensão direta de uma entidade simples.

Para Aristóteles a *substância* é a categoria fundamental, pois é o suporte ou substrato pelo qual a matéria se constitui em algo. Ela é o sujeito de qualquer proposição, é a entidade da qual se diz algo. As demais categorias são utilizadas para se dizer algo sobre a substância, formando o *predicado* de uma proposição (ARISTÓTELES, 1994, p.106).

As categorias aristotélicas possuem duas propriedades lógicas: a **extensão** e a **compreensão**. Denomina-se "extensão" o conjunto de objetos designados por um termo ou uma categoria. A extensão do termo *homem,* por exemplo, será o conjunto de todos os seres que podem ser chamados de homem: Sócrates, Platão, Paulo, Maria. Chama-se "compreensão" o conjunto de propriedades que um termo ou categoria designa. Se dissermos que *homem* é um animal, vertebrado, mamífero, bípede, mortal e racional, essas qualidades formam a compreensão do termo *homem*.

Extensão e compreensão são inversamente proporcionais. Quanto maior a extensão de um termo, menor será a sua compreensão. Quanto maior a compreensão, menor a extensão. Por exemplo, o termo *Jânio Quadros* designa um determinado homem e possui a menor extensão possível, já que se refere a um único ser. Já a sua compreensão possui todas as propriedades do termo *homem*, somada às propriedades específicas de uma determinada pessoa. Essa distinção permite classificar as categorias em três tipos:

- *Gênero*: extensão maior, compreensão menor. Ex.: animal.
- *Espécie*: extensão média e compreensão média. Ex.: homem.

Indivíduo: extensão menor e compreensão maior. Ex.: Jânio Quadros

Em uma proposição, a categoria sustância é o *sujeito* (*S*). As demais categorias formam os *predicados* (*P*). A predicação se faz por meio do verbo de ligação *ser*. Uma proposição é um discurso declarativo que enuncia ou declara verbalmente o que foi pensado e relacionado ao juízo. Ela reúne ou separa verbalmente o que o juízo reuniu ou separou mentalmente. Essa reunião se faz pela afirmação do tipo *S* é *P*. Por exemplo: *Sónia é professora*. A separação se faz pela negação *S* não é *P*, como em *Sonia não é veterinária* (CHAUÍ, 2012, p.139).

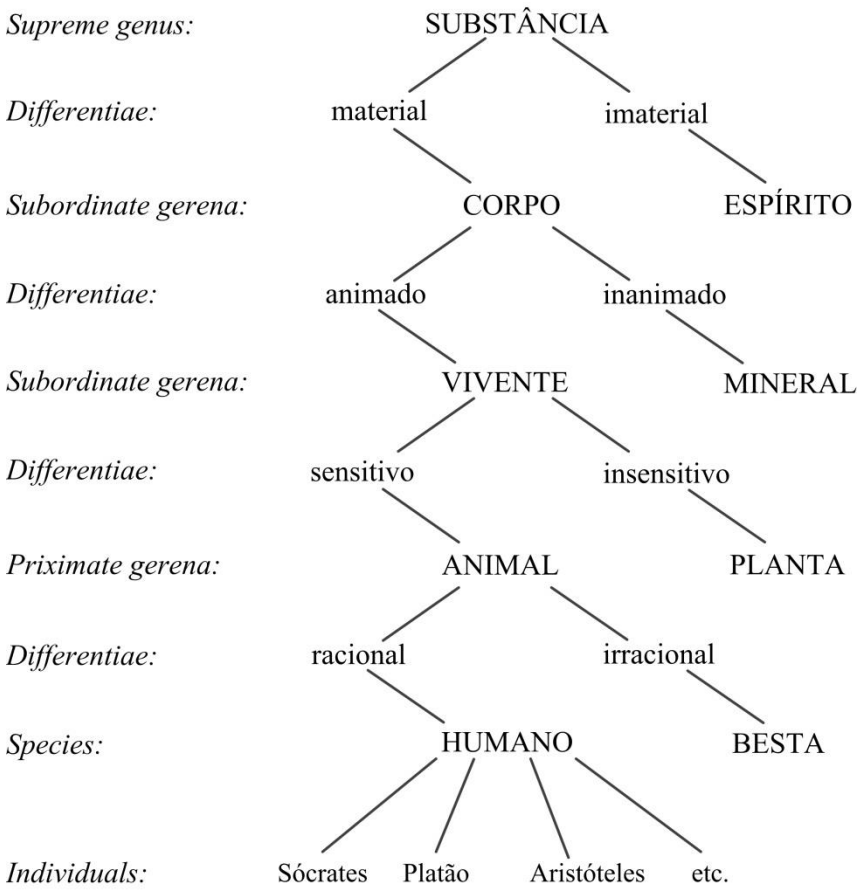
A lógica aristotélica baseia-se na estrutura do sistema de linguagem ocidental, na forma *sujeito-predicado*. O verbo "ser" na maioria dessas línguas tem significado de existência e dele decorrem muitas questões filosóficas.

De todos os grandes pensadores da Grécia antiga, Aristóteles foi o que mais influenciou a civilização ocidental. O modo como pensamos e produzimos conhecimento hoje em dia está direta ou indiretamente relacionado às suas ideias.

Um dos maiores divulgadores das ideias de Aristóteles acerca da categorização foi o filósofo neoplatônico Porfírio (232-304). Dentre as suas contribuições mais importantes está a obra *Introductio in Praedicamenta*, por vezes referenciada pelos títulos de *Isagoge* (título da tradução latina feita pelo filósofo romano Boécio), ou ainda *Quinque voces Porphyrii*. Nesta obra, Porfirio descreve como as qualidades atribuídas às coisas podem ser classificadas, rompendo com o conceito filosófico de substância como uma relação de género/espécie. Dessa forma, pôde incorporar a lógica aristotélica ao neoplatonismo. Porfírio descreve como as qualidades podem ser classificadas e apresentadas em uma estrutura lógica hierárquica que ficou conhecida como "Árvore de Porfírio" (*Arbor porphyriana)*.

A Árvore de Porfírio (Figura 2.5) constitui-se num conjunto hierárquico finito de gêneros e espécies, identificados por dicotomias sucessivas. Embora Porfírio não tenha proposto explicitamente uma representação gráfica, é possível encontrar na literatura variadas representações da estrutura por ele proposta.

Figura 2.5 – Árvore de Porfírio



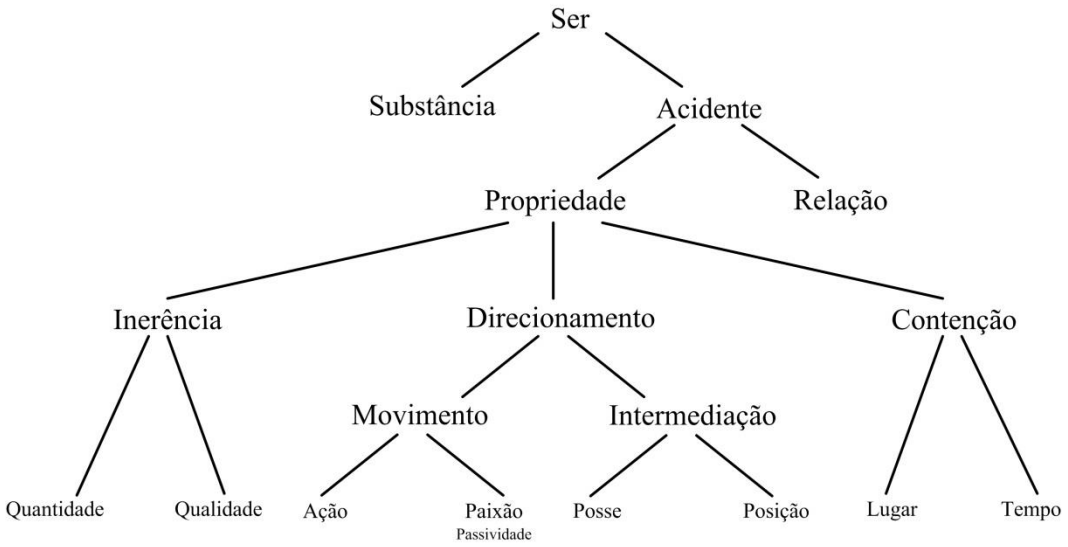
Fonte: Adaptado de Peter of Spain (1239) *apud* Sowa (2000, p.5)

Na Arvore de Porfírio as categorias aristotélicas estão organizadas por *genus* (supertipo) e *species* (subtipo). As características que distingue diferentes *species* do mesmo *genus* são chamadas *differentiae*. Por exemplo: SUBSTÂNCIA com a *differentiae* **material** é CORPO e com *differentiae* **imaterial** é ESPÍRITO. CORPO com *diffetentiae* **animado** é VIVENTE, com *differentiae* **inanimado** é MINERAL (SOWA, 2000, p.4).

Uma das limitações da Árvore de Porfírio é sua dependência de um atributo de qualidade que compreenderá as subdivisões sucessivas. O homem só é mortal numa hierarquia que focalize o problema da duração da vida. A seleção desse atributo de qualidade se dá pelo que Tálamo *et al* (1992) chamam de 'pressão' contextual, podendo considerá-la verdadeira "apenas em relação a um determinado código e não em relação às propriedades dos objetos em si mesmos". Neste tipo de divisão, o contexto não é incorporado, ou então incorpora-se apenas um contexto determinado, o qual irá suportar a interpretação.

O filósofo vienense Franz Brentano (1838-1917) reorganizou as categorias aristotélicas representando-as como folhas de uma árvore cujos galhos são rotulados com termos retirados dos trabalhos de Aristóteles: *Ser*, *Acidente*, *Propriedade*, *Herança*, *Direcionamento*, *Contenção*, *Movimento* e *Intermediação* (SOWA, 2000, p.36).

Figura 2.6 – Árvore de Brentano



Fonte: Adaptado de Sowa (2000, p.57)

Para Brentano a substância deixa de ser a categoria definidora máxima e passa a ser uma instância do *Ser*, que comporta também o *Acidente*; que se subdivide em *Propriedade* e *Relação*. A *Propriedade* está dividida em *Inerência*, *Direcionamento* e *Contenção* em cuja

base estão as demais categorias aristotélicas. Segundo Sowa (2000, p.57), esse arranjo mais sistematizado das categorias aristotélicas é fundamental para todas as futuras formas de representação do conhecimento.

Para Aristóteles, a essência das coisas é exclusivamente determinada pelas próprias coisas. Immanuel Kant (1724-1804), seguindo as ideias do filósofo e historiador escocês David Hume (1711-1776), defende em sua obra "Crítica da Razão Pura" que a essência de uma coisa não poderia ser separada de quem a percebe. Segundo Chauí (2012, p.249):

> [...] Kant realizou uma 'revolução copernicana' em filosofia, isto é, exigiu que, antes de qualquer afirmação sobre as ideias, houvesse o estudo da própria capacidade de conhecer, isto é, da razão, e que era preciso mostrar que a razão não depende das coisas nem é regulada por elas e sim as coisas dependem da razão e são reguladas por ela.

Ao estabelecer os critérios para o conhecimento, Kant propõe uma reformulação das categorias aristotélicas, buscando torná-las dependentes do "juízo", sendo que um juízo é a afirmação ou a negação da realidade de um objeto pela afirmação ou negação de suas propriedades. Segundo Chauí (2012, p.250):

> Um juízo, portanto, nos dá a conhecer alguma coisa, desde que esta possa ser apreendida sob as formas do espaço e do tempo e sob os conceitos do entendimento. Uma coisa passa a existir quando se torna objeto de um juízo. Isso não significa que o juízo cria a própria coisa, mas sim que a faz existir para nós. O juízo põe a realidade de alguma coisa ao colocá-la como sujeito de uma proposição, isto é, ao colocá-la como objeto de um conhecimento. É, portanto, o juízo que põe a qualidade, a quantidade, a causalidade, a substância, a matéria, a forma, a essência das coisas, na medida em que estas existem apenas enquanto são objetos de conhecimento postos pelas formas do espaço, do tempo e pelos conceitos do entendimento.

A partir de sua *tábua dos juízos*, Kant propõe uma correspondente *tábua das categorias*, ambas organizadas em quatro grupos contendo três elementos (Quadro 2.2). A classificação lógica dos juízos é o fio condutor para as categorias. Para cada espécie de juízo pode-se abstrair um conceito máximo, uma categoria. Kant chama esse procedimento de "dedução metafísica das categorias".

Quadro 2.2 – Tabua dos Juízos e Categorias de Kant

	JUÍZOS	CATEGORIAS
Quantidade	Universais Particulares Singulares	Unidade Pluralidade Totalidade
Qualidade	Afirmativos Negativos Infinitos	Realidade Negação Limitação
Relação	Categóricos Hipotéticos Disjuntivos	Inerência Causalidade Comunidade
Modalidade	Problemáticos Assertóricos Apodíticos	Possibilidade Existência Necessidade

Fonte: Adaptado de Salatiel (2006, p. 82).

As sensações envolvidas quando uma pessoa percebe a realidade são postas em ordem primeiro no espaço e no tempo, e então, de acordo com estas categorias. Segundo Kant, o juízo dá-se quando superpomos categorias à associação dos conceitos que constituem a proposição. Por exemplo, na proposição "essa rosa é vermelha" temos um juízo *singular*, *afirmativo*, *categórico* e *assertórico*. Correspondentemente, as categorias aplicadas são: *totalidade*, *realidade*, *inerência* e *existência*. Na proposição "se um metal é aquecido ele se expande" temos um juízo *universal*, concernente a todos os metais, ele é *afirmativo*, é *hipotético* e *assertórico*. As categorias aplicadas são respectivamente de *unidade*, *realidade*, *causalidade* e *existência*.

Segundo Kant, "vemos o mundo através das nossas lentes cognitivas". Sabe-se hoje que essas "lentes" não são exatamente como Kant imaginou e, certamente não são iguais para todos os humanos, pois dependem de fatores como cultura e meio social. Entretanto, um dos legados de Kant é a consciência de que o conhecimento não é "um espelho da natureza" e não se dá apenas pelo acúmulo de percepções ou observações; ele depende da criatividade, da imaginação e do poder de abstração do nosso intelecto (SILVEIRA, 2002).

Charles Sandres Peirce (1839-1914) observou que algumas tríades de Kant refletiam três categorias mais básicas, que foram denominadas de Primeiridade (*firstness*), Secundidade (*secondness*) e Terceiridade (*thirdness*). Segundo Sowa (2000, p.61), a *Primeiridade* é determinada pelas qualidades inerentes em alguma coisa. É a concepção do Ser independente de qualquer outra coisa. A *Secundidade* é determinada pela relação ou reação direcionada a outra coisa. É a concepção do Ser relativo a ou reativo a alguma coisa. A *Terceiridade* é

determinada por alguma mediação que agrega várias entidades em uma relação. É a concepção de mediação, em que um primeiro e um segundo são postos em relação.

Sowa (2000, p.61) exemplifica as categorias peircenianas de forma bem ilustrativa. Um indivíduo pode ser reconhecido como um ser humano ou como um subtipo, tal como homem ou mulher, por meio de impressões sensoriais, independente de quaisquer relações externas. O tipo rotulado como Mulher caracteriza um indivíduo por meio de propriedades que podem ser reconhecidas sem considerar quaisquer relacionamentos com outras entidades (Primeiridade).

O mesmo indivíduo poderia ser classificado relativamente a muitas outras coisas, como no conceito de *mãe*, *advogado*, *esposa*, *piloto*, *empregado* ou *pedestre*. Essa classificação depende de um relacionamento externo com alguma outra entidade, tal como *criança*, *cliente*, *marido*, *avião*, *empregador* ou *tráfego* (Secundidade)

A Terceiridade se foca na mediação que trás a primeira e a segunda na relação. Maternidade, que compõe o ato de dar a luz e o período subsequente de amamentação, relaciona a mãe e a criança. O sistema jurídico dá origem aos papéis de advogado e cliente. O casamento relaciona a esposa e o marido. A aviação relaciona o piloto ao avião. A empresa comercial relaciona o empregado ao empregador. E a atividade de caminhar na rua que é dominada por veículos relaciona o pedestre e o tráfego.

Com o trabalho de Peirce tomou-se consciência da linguagem em sentido amplo, o que gerou a necessidade do aparecimento de uma ciência capaz de criar dispositivos de indagação e instrumentos metodológicos aptos a desvendar o universo multiforme e diversificado dos fenômenos de linguagem (SANTAELLA, 1994, p.15).

Os filósofos e pensadores aqui apresentados, assim como todo ser humano, cada qual conforme o aparato intelectual de que dispõe, querem compreender o mundo buscando sua ordem, sua estabilidade, ainda que reconheçam a sua inerente dinamicidade.

2.2 Ontologia na Ciência da Computação

A bibliografia da área aponta que a primeira menção do termo ontologia em um trabalho de Ciência da Computação se deu no artigo intitulado *Another look at data*, de George H. Mealy (1967). Desde então o tema "ontologia" têm 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*)[3], 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 na área é a de Gruber (1993), que descreve uma ontologia como: "uma especificação explícita de uma conceitualização". Porém, a partir da leitura de um artigo posterior do mesmo autor, pode-se, por fim, definir 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 como 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:

> 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).

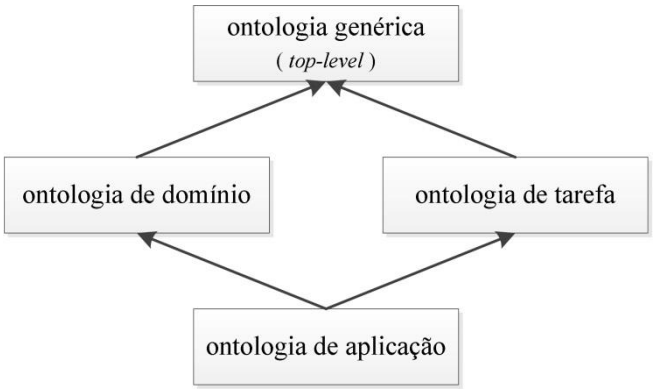
Jasper e Uschold (1999) ressaltam a necessidade de explicitação dos relacionamentos entre os termos de uma ontologia:

> 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.

[3] http://www.formalontology.org/

Guarino (1998) propõe que ontologias sejam construídas segundo seu nível de generalidade, de modo que os conceitos de uma ontologia de domínio ou de tarefa sejam especializações dos termos de uma ontologia genérica (Figura 2.7).

Figura 2.7 – Tipos de ontologias



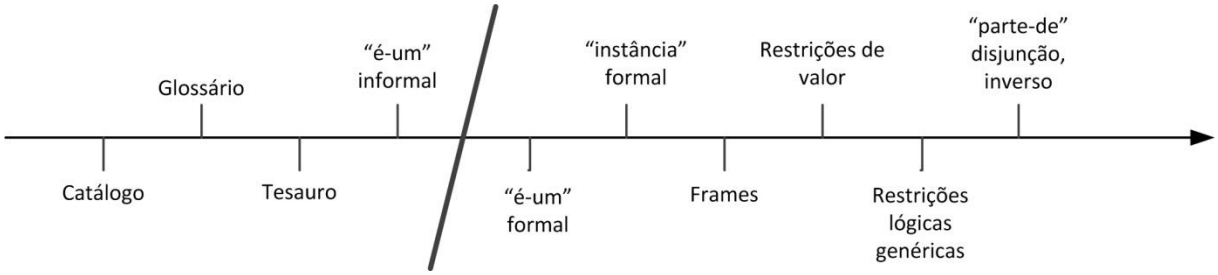
Fonte: Adaptado de Guarino (1998)

Os conceitos de uma ontologia de aplicação, por sua vez, devem ser especializações dos termos das ontologias de domínio e de tarefa correspondentes. Pode-se descrever sucintamente cada uma destes tipos de ontologias da seguinte forma:

- **ontologias genéricas (*top-level ontologies*)**: descrevem conceitos gerais como espaço, tempo, matéria, objeto, evento, ação, etc., que são independentes de um problema ou domínio particular;
- **ontologias de domínio** e **ontologias de tarefa**: descrevem, respectivamente, o vocabulário relacionado a um domínio como medicina, direito, etc, ou uma tarefa ou atividade como diagnóstico, vendas, etc, especializando os termos introduzidos no nível superior
- **ontologias de aplicação:** descrevem conceitos dependentes de um determinado domínio e tarefa particular. Estes conceitos frequentemente correspondem aos papéis desempenhados por entidades do domínio ao executar uma determinada atividade.

Com o propósito de classificar as ontologias de acordo com o seu nível de expressividade, Lassila e MacGuinness (2003) apresentam um esquema ("Espectro Ontológico") que abrange desde instrumentos com um baixo nível de expressividade semântica até instrumentos que possibilitam definir relações mais complexas (Figura 2.8).

Figura 2.8 – Espectro ontológico



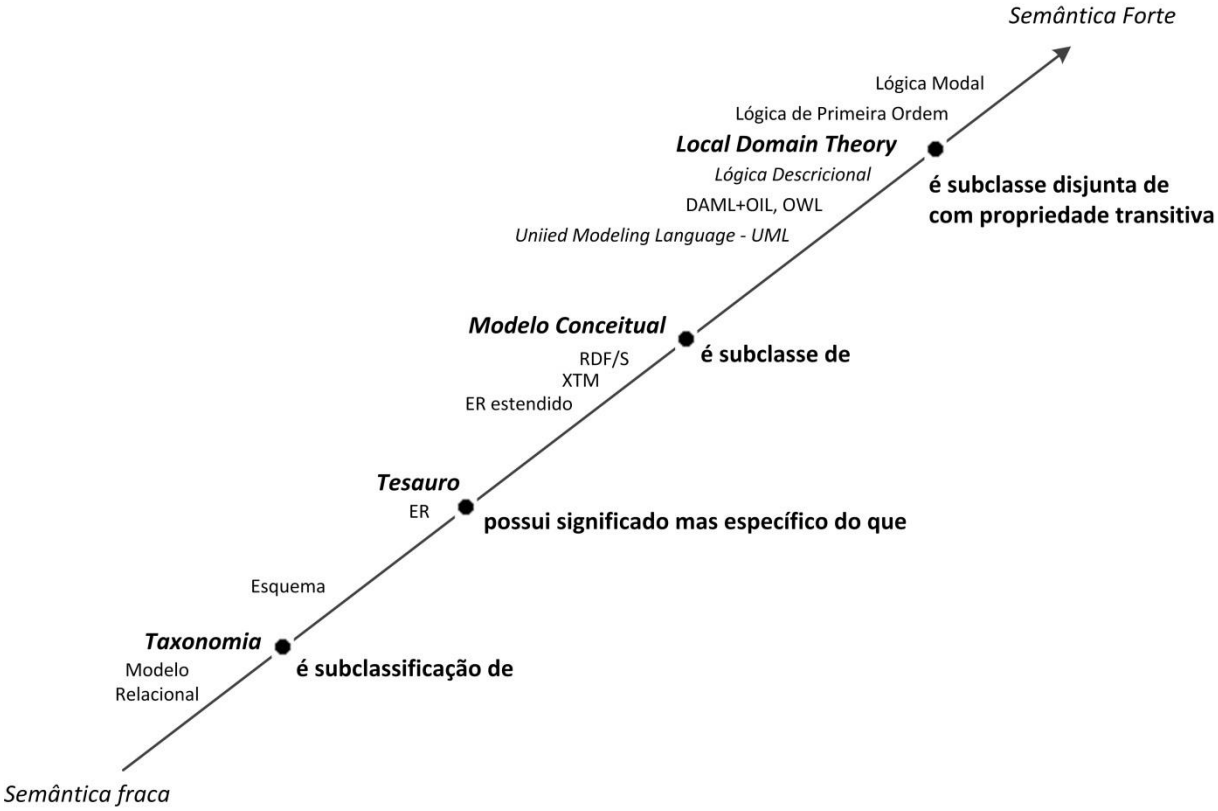
Fonte: Adaptado de Lassila e McGuinness (2001)

- **Catálogo**: lista de termos e acrônimos de um determinado domínio e suas respectivas interpretações.
- **Glossário**: lista de termos com seus respectivos significados especificados em linguagem natural.
- **Tesauros**: vocabulário controlado que incorpora semântica por meio das relações entre termos. Embora não apresente explicitamente uma forma hierarquia, esta pode ser deduzida por meio das relações de generalização e especialização.
- **"é-um" informal**: hierarquias que utilizam relacionamentos informais, de modo que conceitos podem ser livremente associados a uma categoria, mesmo que formalmente não faça parte da mesma.
- **"é-um" formal**: incluem relações estritas de subclasse, que permitem incorporar o conceito de herança.
- **"instância" formal**: utilizam formalmente relacionamentos classe-instância;
- **Frames**: as classes possuem propriedades específicas, possibilitando contextualizar informações em um determinado domínio, sendo herdadas por subclasses e instâncias;
- **Restrições de valor:** definem restrições para os valores assumidos pelas propriedades;
- **Restrições lógicas genéricas**: possibilitam a definição de restrições lógicas utilizando valores de propriedades de diversas classes ou instâncias;
- **"Parte-de", disjunção, inverso**: permitem especificar classes disjuntas, relacionamentos inversos e relacionamentos do tipo parte-todo.

A barra inclinada na Figura 2.8 separa os instrumentos comumente utilizados por humanos daqueles descritos em linguagens formais, desenvolvidos para serem utilizados em ambientes computacionais.

Da mesma forma, Daconta, Obrst e Smith (2003, p.157) definem também um "Espectro Ontológico" (Figura 2.9) que apresenta conceitos relacionados a questões de representação, classificação e desambiguação semântica. Segundo estes autores, o que é normalmente conhecido como ontologia pode variar da simples noção de uma taxonomia até um modelo lógico, passando por um tesauro, um modelo conceptual, a lógica descricional, a lógica de primeira ordem, etc.

Figura 2.9 – Espectro ontológico: da semântica fraca para a semântica forte



Fonte: Adaptado de Daconta, Obrst e Smith (2003, p.157)

Em uma *taxionomia* a semântica das relações entre um nó-pai e um nó-filho não existe ou é relativamente mal definida. Em alguns casos subentende-se a relação *subclasse de*; em outros se supõe a relação *parte de*. Ainda em alguns outros casos a relação pai-filho pode ser indefinida. Em um *tesauro*, um determinado vocabulário está organizado em uma ordem preestabelecida e estruturado de modo que os relacionamentos de equivalência, de homografia, de hierarquia e de associação entre termos sejam indicados claramente e

identificados por indicadores padronizados. Já um modelo conceitual: é uma representação de um domínio; suas entidades e as relações entre elas; seus atributos e seus valores; e as regras que associam entidades, relacionamentos e atributos.

Como pode ser observado, esses dois "espectros ontológicos" não utilizam nem fazem referência à ontologia como um objeto autônomo. Pode-se concluir, assim, que em certo nível de abstração uma ontologia refere-se à ideia de representatividade ou expressividade semântica. O modelo relacional, o modelo entidade-relacionamento, a linguagem UML são ferramentas que se apresentam em um grau crescente de seu poder de representar a realidade de um domínio. Se tais ferramentas podem ou não serem aceitas ou definidas como ontologias, depende de como se define "ontologia" (CORCHO; FERNÁNDEZ-LÓPEZ; GÓMEZ-PÈREZ, 2003).

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). 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 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.

Para a Ciência da Computação uma ontologia é, enfim, uma estrutura conceitual que visa representar formalmente os conceitos e suas relações, regras e restrições lógicas de um determinado domínio. Pode ser definida por meio de linguagens legíveis e processáveis por computadores.

2.3 Ontologia e a Ciência da Informação

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.

Como já visto anteriormente, a organização do conhecimento foi uma preocupação de muitos pensadores, tais como: Aristóteles, Kant, Peirce, entre outros. No entanto, foi apenas no final do século XIX que a organização do conhecimento consolidou-se como área que visa a gestão do conhecimento contido em documentos, com o desenvolvimento de sistemas de conhecimento geral, como a *Classificação Decimal Dewey* e a *Classificação Decimal Universal* (GNOLI, 2009).

De forma contrária à ideia por trás dos conceitos de organização e representação do conhecimento, Dahlberg (2006) diz que, em essência, todo conhecimento possui uma natureza subjetiva, individual e não é transferível, somente podendo ser adquirido por meio da reflexão. Contudo, Fujita (2008) argumenta que tal "conhecimento subjetivo e individual poderá ser transferido mediante formas de representação escrita ou falada, considerando-se nosso conhecimento prévio linguístico que expressará nossas experiências e compreensões".

Considerando apenas os elementos básicos de uma ontologia, os conceitos e as relações entre eles, algumas das tradicionais habilidades do profissional da informação – construção de vocabulários controlados, concepção de classificações/taxionomia, organização da informação – contribuem significativamente na tarefa de elaborar ontologias. Estas habilidades estão apoiadas em sólidas bases teóricas e metodológicas, desenvolvidas ao longo de décadas em obras como as de Otlet (1934), Ranganathan (1967), Dahlberg (1978), Hjørland (2002), entre outras. Porém, uma ontologia é mais que um vocabulário controlado, uma taxonomia ou um tesauro. É uma estrutura sistêmica que possui mecanismos (regras, axiomas, etc.) que permitem realizar restrições lógicas e inferências sobre os seus conceitos e relações.

Se na Computação uma ontologia pode ser considerada como um artefato tecnológico, no contexto da Ciência da Informação caracteriza-se como um instrumento de

nível epistemológico, concebido para favorecer a representação formal dos conceitos e dos relacionamentos existentes entre eles em um domínio específico.

Segundo Ramalho:

> 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 (2010, p.37).

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).

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.

Assim como a Web Semântica, ontologia é tema de pesquisa comum entre a Ciência da Computação e a Ciência da Informação. Cabe a essas ciências estreitarem suas relações e promover um diálogo mais constante e efetivo para um desenvolvimento rápido e sólido de ambas.

2.4 Resumo e Discussão

Este capítulo apresentou inicialmente um histórico da utilização do termo "ontologia", desde a primeira menção na obra de Jacob Lorhard, em 1606, até a consolidação do termo na Filosofia a partir da obra de Christian Wolff, em 1730. Em seguida foi

apresentada a etimologia da palavra e a sua definição em alguns dicionários da língua portuguesa.

Primariamente referindo-se a uma disciplina da Filosofia, as ontologias vêm ocupando cada vez mais a atenção de outras áreas, sendo utilizadas de diferentes maneiras. Na Ciência da Computação é vista de forma pragmática, constituindo de uma estrutura conceitual que tem por objetivo representar os elementos ou o conhecimento de um determinado domínio. Já na área da Ciência da Informação, as ontologias vêm se somar a outras ferramentas de representação e organização da informação, que há décadas vêm sendo estudadas e utilizadas.

As ontologias se configuram como um tema de fronteira entre A Ciência da Informação e a Ciência da Computação. É desejável que futuras pesquisas venham a ser desenvolvidas de forma integrada, buscando trazer à Ciência da Informação conhecimentos e ideias da Ciência da Computação. Da mesma forma, as pesquisas em Ciência da Computação devem considerar a existência de uma ciência que há muito tempo vem abordando de forma sistemática os problemas relacionados à classificação, representação e recuperação de informação.

Referências

ARISTÓTELES. **Categorias de Aristóteles**. Tradução de Silvestre Pinheiro Ferreira. 3ª ed. Lisboa: Guimarães Editores, 1994.

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.

BUSH, V. As We May Think. **The Atlantic Monthly**, July, 1945.

CASTRO, S. **Ontologia**. Rio de Janeiro: Jorge Zahar, 2008.

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.

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

CORCHO, O.; FERNÁNDEZ-LÓPEZ , M, GÓMEZ-PÉREZ, M. Methodologies, tools and languages for building ontologies. Where is their meeting point? **Data & Knowledge Engineering**, v.46, n.1, 2003.

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.

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

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.

HJØRLAND, B. Domain analysis in information science: Eleven approaches - traditional as well as innovative. **Journal of Documentation**, v. 58, n. 4, 2002.

JACOB, E. K. Ontologies and the semantic web. **Bulletin of the American Society for Information Science and Technology**, apr./may,, 2003.

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

GARCÍA MARCO, F. J. Ontologías y organización del conocimiento: retos y oportunidades para el profesional de la información. **El profesional de la información**, v.16, n.6, 2007.

GNOLI, C. The ontological approach to knowledge organization**. Invited paper sent for discussion at the 2 Seminário de pesquisa em Ontologia no Brasil.** Rio de Janeiro, Setembro, 2009.

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.

GRUBER, T. **What is an Ontology?** 1992. Disponível em: <http://wwwksl. stanford.edu/kst/what-is-an-ontology.html>. Acesso em: 18/01/2013.

GUIZZARDI, G. **Ontological Foundations for Structural Conceptual Models**. The Netherlands: Universal Press, 2005. Disponível em <http://doc.utwente.nl/50826/1/thesis_Guizzardi.pdf>. Acesso em 14.02.2013.

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

LASSILA, O.; McGUINNESS, D.L. **The Role of Frame-Based Representation on the Semantic Web**. Knowledge Systems Laboratory Report KSL-01-02, Stanford University, 2001.

LIMA-MARQUES, M. **Ontologias: da filosofia à representação do conhecimento**. Brasília: Thesausus, 2006.

McGUINNESS, D.L. Ontologies come of age. In: FENSEL, D., HENDLER, J., LIEBERMAN, H., WAHLSTER, W. (Eds.), **Spinning the Semantic Web**: Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge, MA, 2003, pp. 171–196.

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

MOREIRA, W. **A construção de informações documentárias: aportes da linguística documentária, da terminologia e das ontologias**. Tese (Doutorado em Ciência da Informação) ─ Universidade de São Paulo, Escola de Comunicação e Artes, 2010.

NOVELLO, T.C. **Ontologias**: sistemas baseados em conhecimento e modelos de banco de dados. Universidade Federal do Rio Grande do Sul, 2002.

OTLET, P. **Traité de documentation**: le livre sur le livre, théorie et pratique. Bruxelles: Editiones Mundaneum, 1934.

ØHRSTRØM, P.; SCHÄRFE, H.; UCKELMAN, S.L. Jacob Lorhard's ontology: a 17th century hypertext on the reality and temporality of the world of intelligibles, In: **Proceedings of the 16th International Conference on Conceptual Structures** (ICCS). Berlin: Springer-Verlag, 2008.

PORFÍRIO DE TIRO. **Isagoge**: introdução às *Categorias* de Aristóteles. Introdução, tradução e comentários de Bento Silva Santos. São Paulo: Attar Editorial, 2002.

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.

RANGANATHAN, S. R. **Prolegomena to library classification**. New York: AsiaPublishing House, 1967.

SMITH, B. **Ontology and information systems**. 2002. Disponível em: <http://ontology.buffalo.edu/ontology(PIC).pdf>. Acesso em: 17 fev. 2014.

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

SOWA, J.F. **Knowledge Representation**: Logical, Philosophical, and Computacional Foundations. Pacific Grove: Brooks/Cole, 2000.

SALATIEL, J. R. Peirce e Kant sobre categorias: parte I: dedução metafísica e reviravolta semiótica. **Cognitio-Estudos: Revista Eletrônica de Filosofia**, v. 3, n. 1, p. 79-88, jan./jun. 2006.

SANTAELLA, L. **O que é Semiótica**. 11ª ed. São Paulo: Brasiliense, 1994. Coleção Primeiros Passos – 103.

SILVEIRA, F.L. A Teoria do Conhecimento de Kant: o Idealismo transcendental. **Caderno Brasileiro de Ensino de Física**. v.19, número especial, jun. 2002.

TÁLAMO, M.F.G.M. *et al* Instrumentos de controle terminológico: limites e funções. In: **Anais do II Simpósio Lationamericano de Terminologia**. Brasília: União Latina/ IBICT, 1992.

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

3
Recuperação de Informação baseada em Ontologia

Este capítulo apresenta a origem do termo "Recuperação de Informação", que se estabeleceu como uma área de pesquisa de interesse comum entre a Ciência da Computação e a Ciência da Informação. Descreve-se resumidamente o processo de recuperação de informação e seus principais modelos. Por fim apresenta formas de inserção das ontologias na recuperação de informação e propõe critérios de classificação de sistemas baseados em ontologia.

Em 1951, Calvin Mooers criou o termo "*Information Retrieval*" (Recuperação de Informação) e definiu os problemas a serem abordados por esta nova disciplina.

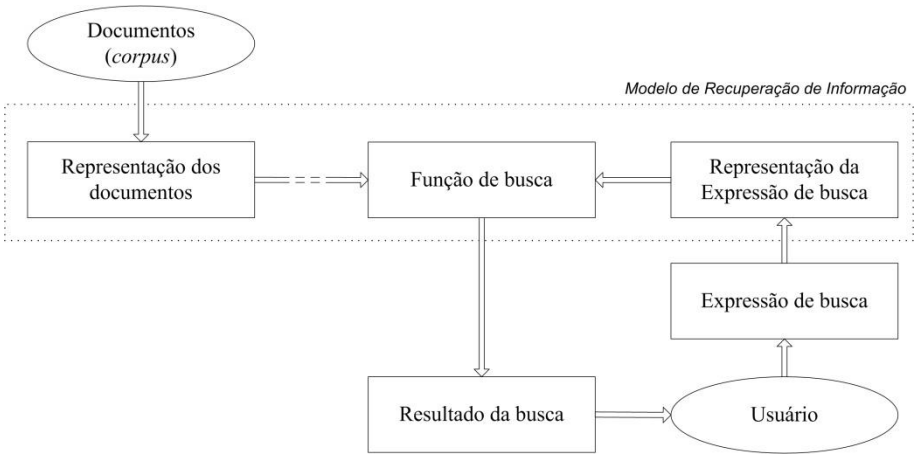
> [A Recuperação de Informação] trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação (MOOERS, 1951).

Para Saracevic (1999), a Recuperação de Informação pode ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação.

Na Ciência da Computação a Recuperação de Informação (*Information Retrieval*) se firmou como uma área de pesquisa autônoma cujo interesse está centrado no desenvolvimento de ferramentas para o tratamento de fontes de informação não estruturadas e semiestruturadas. É tema de interesse de uma imensa comunidade de pesquisadores de todas as partes do mundo e abriga uma grande quantidade de vertentes, abordagens e metodologias para os problemas dessa área.

Recuperar uma informação consiste em identificar, em um acervo documental, quais aqueles que satisfazem total ou parcialmente a uma determinada necessidade de informação do usuário. Considera-se, em princípio, que o usuário está interessado em recuperar "informação" sobre um determinado assunto e não documentos, embora seja nestes que a informação está registrada. Uma representação simplificada do processo de recuperação de informação é apresentada na Figura 3.1.

Figura 3.1 – Representação do processo de recuperação de informação



Fonte: Adaptado de FERNEDA, 2012, p.14

Buckland (1991) define o conceito de "informação como coisa*"* e argumenta que os acervos dos sistemas de informação seriam registros relacionados a coisas ou objetos. Nesses sistemas informação está vinculada ao objeto que a contém. Por sua vez, o termo documento, entendido como coisa informativa, incluiria objetos, artefatos, imagens, sons etc.

Independente dos tipos de documentos gerenciados por um sistema de informação, a eficiência da recuperação é dependente da forma como esses documentos estão representados. A representação de um documento tem por objetivo identificar e descrever resumidamente o seu conteúdo informacional, ao mesmo tempo em que define seus pontos de acesso para a busca em um sistema de informação. A tarefa de representar os documentos é feita em um tempo anterior à execução de qualquer busca. No esquema da Figura 3.1, a existência de uma seta seccionada tenta mostrar essa assincronia.

Em um sistema de recuperação de informação o usuário expressa sua necessidade de informação por meio de uma expressão de busca, composta geralmente por um conjunto de termos que representa linguisticamente a necessidade de informação do usuário. A principal dificuldade está em predizer os termos foram usados para representar os documentos que

satisfarão sua necessidade, e ao mesmo tempo evitar a recuperação de documentos não relevantes. Para isso é necessário que o usuário tenha um relativo conhecimento do vocabulário ou da terminologia do domínio de conhecimento de seu interesse.

Para que seja possível uma comparação entre a expressão de busca e cada um dos documentos do *corpus* é necessário que essa seja representada de forma similar à representação dos documentos. Diversos recursos podem ser oferecidos por um sistema a fim de facilitar o usuário na especificação de sua expressão de busca. Porém, essa consulta deve ser representada de uma forma idêntica ou similar à utilizada pelos documentos.

No centro do processo de recuperação de informação está a função de busca, que compara as representações dos documentos com a representação da expressão de busca e recupera os itens que supostamente fornecerão informações relevantes. De forma geral, a função de busca calcula o grau de similaridade entre a expressão de busca e cada um dos documentos do *corpus*. Na maioria dos sistemas, a similaridade é definida por um valor numérico que pretensamente define o quão relevante é um documento para satisfazer a necessidade de informação do usuário. O resultado de uma busca é geralmente composto por uma lista de referências a documentos, ordenada pelo grau de similaridade calculada pela função de busca.

Um modelo de recuperação de informação é a especificação formal de três elementos: a *representação dos documentos*, a *representação da expressão de busca* e a *função de busca* (FERNEDA, 2012, p.20). De maneira mais formal, Baeza-Yates e Ribeiro-Neto (2011, p.58) definem modelo de recuperação de informação como uma quadrupla:

[**D**, **Q**, F, $R(q_i, d_j)$]

D é um conjunto composto por visões lógicas (representações) dos documentos no *corpus*;

Q é um conjunto composto de visões lógicas das necessidades de informação dos usuários;

F é um *framework* para a modelagem de representações dos documentos, consultas e seus relacionamentos;

$R(q_i, d_j)$ é uma função de ordenamento (*ranking*) que atribui um número real à relação entre uma representação da consulta q_i de **Q** e a representação de um documento d_j de **D**.

Apesar de alguns dos modelos de recuperação de informação terem sido criados entre as décadas de 1960 e 1970, as suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da Web.

3.1 Modelos de Recuperação de Informação

Nessa seção serão descritos os chamado "modelos clássicos" de recuperação de informação, cujas propostas serviram de base para o desenvolvimento de diversos outros modelos e algumas técnicas que até hoje são utilizadas. Os modelos clássicos são: modelo booleano, modelo espaço vetorial e o modelo probabilístico.

3.1.1 Modelo Booleano

No Modelo Booleano (FERNEDA, 2012, cap.3) um documento é representado por um conjunto de termos de indexação. As buscas são formuladas por meio de uma expressão booleana composta por termos ligados através dos operadores lógicos AND, OR e NOT, e apresentam como resultado o conjunto de documentos cuja representação satisfaz as restrições lógicas da expressão de busca.

Uma expressão conjuntiva de enunciado $\boldsymbol{t_1}$ **AND** $\boldsymbol{t_2}$ recuperará documentos indexados por ambos os termos ($\boldsymbol{t_1}$ e $\boldsymbol{t_2}$). Uma expressão disjuntiva $\boldsymbol{t_1}$ **OR** $\boldsymbol{t_2}$ recuperará o conjunto dos documentos indexados pelo termo $\boldsymbol{t_1}$ ou pelo termo $\boldsymbol{t_2}$. Uma expressão que utiliza apenas um termo $\boldsymbol{t_1}$ terá como resultado o conjunto de documentos indexados por esse termo. A expressão **NOT** $\boldsymbol{t_1}$ recuperará os documentos que não são indexados pelo termo $\boldsymbol{t_1}$. As expressões $\boldsymbol{t_1}$ **NOT** $\boldsymbol{t_2}$ ou $\boldsymbol{t_1}$ **AND NOT** $\boldsymbol{t_2}$ terão como resultado o conjunto dos documentos que são indexados por $\boldsymbol{t_1}$ e que não são indexados por $\boldsymbol{t_2}$.

Termos e operadores booleanos podem ser combinados para especificar buscas mais detalhadas ou restritivas. Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem delimitando partes da expressão por meio de parênteses. A definição de expressões complexas exige um

conhecimento profundo da lógica booleana. O conhecimento da lógica booleana é importante também para entender e avaliar os resultados obtidos em uma busca.

O modelo booleano possui diversas limitações. Algumas delas são:

- Não existe uma forma de atribuir importância relativa (pesos) aos termos de indexação dos documentos nem aos diferentes termos da expressão de busca. Assume-se implicitamente que todos os termos possuem o mesmo peso.
- O resultado de uma busca booleana se caracteriza por uma simples partição do *corpus* em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual relevância, não havendo nenhum mecanismo pelo qual os documentos possam ser ordenados;
- Sem um treinamento apropriado, o usuário leigo será capaz de formular somente buscas simples. Para buscas que exijam expressões mais complexas é necessário um conhecimento sólido da lógica booleana.

Apesar de suas limitações, muitos sistemas se desenvolveram utilizando o modelo booleano como ponto de partida para a implementação de novos recursos de recuperação. Neste sentido o modelo booleano pode ser considerado o modelo mais utilizado nos sistemas de recuperação de informação e nos mecanismos de busca da Web.

3.1.2 Modelo Espaço Vetorial

No Modelo Espaço Vetorial (ou simplesmente Modelo Vetorial), um documento é representado por um vetor numérico onde cada elemento representa o peso, ou relevância, do respectivo termo de indexação na representação do conteúdo informacional do documento. Da mesma forma que os documentos, uma expressão de busca também é representada por um vetor numérico onde cada elemento representa a importância (peso) do respectivo termo na representação da necessidade de informação do usuário (SALTON; WONG; YANG, 1975).

A utilização de uma mesma representação tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade entre o vetor que representa uma determinada expressão de busca e cada um dos vetores que representam os documentos do *corpus*. Em um espaço vetorial contendo $\boldsymbol{N}$ dimensões, a similaridade ($\boldsymbol{sim}$) entre um

documento $\boldsymbol{d_j}$ e uma expressão de busca $\boldsymbol{q}$ é obtida por meio do da seguinte fórmula (SALTON; McGILL, 1983, p.121):

$$sim(d_j, q) = \frac{\sum_{i=1}^{N} (w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{N} w^2_{i,q}}}$$

onde $w_{i,j}$ é o peso do *i*-ésimo termo do documento d_j e $w_{i,q}$ é o peso do *i*-ésimo termo da expressão de busca *q*.

Os valores da similaridade (*sim*) entre uma expressão de busca e cada um dos documentos do *corpus* são utilizados no ordenamento dos documentos resultantes. Portanto, no modelo vetorial o resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade entre cada documento e a expressão de busca.

Um dos maiores méritos do modelo vetorial é a definição de um dos componentes essenciais de qualquer teoria científica: um modelo conceitual. Este modelo serviu como base para o desenvolvimento de uma teoria que alimentou uma grande quantidade de pesquisas e resultou em sistemas como o SMART (SALTON, 1971).

3.1.3 Modelo Probabilístico

O Modelo Probabilístico foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores, tais como Robertson e Jones (1976). A ideia é tratar o processo de recuperação de informação como um processo probabilístico, já que é caracterizado por seu grau de incerteza no julgamento de relevância dos documentos em relação a uma expressão de busca. Assim, é mais realista pensar em uma probabilidade de relevância do que em uma pretensa relevância exata, como a utilizada nos modelos booleano e vetorial.

A partir de uma expressão de busca, composta por um ou mais termos, o usuário expressa sua necessidade de informação e a submete ao sistema. Por meio de cálculos de probabilidade o sistema calcula, para cada documento do *corpus*, um valor numérico (similaridade), que representa a provável relevância do documento para a consulta. Esse valor é utilizado para ordenar os resultados da busca. Tendo um primeiro conjunto de documentos, o usuário pode marca alguns deles que considera verdadeiramente relevantes para a sua necessidade. O conjunto de documentos marcados pode ser então submetido ao sistema,

permitindo fornecer resultados mais precisos. Esse processo, denominado *relevance feedback*, pode ser repetido até que o usuário se sinta satisfeito com os resultados.

Uma virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa do usuário, pois é o único modelo que incorpora explicitamente o processo de *relevance feedback* como base para a sua operacionalização.

Os três modelos clássicos compartilham um mesmo paradigma de representação no qual os termos extraídos dos documentos e das buscas são suficientes para se efetivar o processo de recuperação de informação, considerando-o um sistema fechado no qual o significado dos elementos lexicais é dado pelas suas inter-relações no interior de um *corpus* documental.

A ideia de agregar um conhecimento externo ao processo de recuperação de informação tem por objetivo enriquecer as representações dos documentos e das buscas e compatibiliza-las, proporcionando uma melhoria nos resultados. Neste trabalho, as ontologias são consideradas uma forma de vocabulário controlado cujos termos (conceitos) podem ser inseridos nas representações dos documentos e nas buscas sobre um determinado domínio ou assunto, padronizando a terminologia utilizada em tais representações.

3.2 Ontologia na Recuperação de Informação

Considerando o processo de recuperação de informação (Figura 3.1), observa-se que existem dois elementos de representação que afetam diretamente na eficiência da recuperação de informação: a *representação dos documentos* e a *representação da expressão de busca*. Esses dois elementos se relacionam com entidades externas ao *modelo de recuperação de informação*, respectivamente aos *documentos (corpus)* e ao *usuário*. Esses dois elementos, posicionados em lados extremos do processo de recuperação, possuem natureza linguística. Assim, a eficiência do processo de recuperação de informação é dependente da correta interpretação e representação desses dois elementos linguísticos. 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. Um maior detalhamento sobre a indexação automática será apresentada no Capítulo 4.

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. Esse processo, denominado "expansão de consulta" é detalhado no Capítulo 5.

A partir da literatura da área da Ciência da Computação é possível elencar alguns processos relacionados à recuperação de informação onde as ontologias atualmente estão sendo utilizadas:

- **Indexação automática baseada em ontologia**: o índice que representa um documento é acrescido de termos automaticamente derivados de uma ontologia;
- **Expansão de consulta baseada em ontologia:** a consulta do usuário é modificada com a adição de conceitos provenientes de uma ontologia;
- **Sistemas de recuperação de informação semânticos**: os documentos são previamente anotados (marcados) de acordo com uma ontologia de domínio;
- **Sistemas de coleta de informação baseados em ontologia**: desempenham funções de processamento de textos, tais como classificação, extração e busca;
- **Interfaces de busca**: os conceitos de uma ontologia são apresentados ao usuário que seleciona aqueles que serão utilizados como termos de busca.

O modelo proposto neste trabalho utiliza o Modelo Vetorial (Seção 3.1.2) como arquitetura básica. Ontologias são utilizadas como suporte à indexação automática e como

recurso auxiliar na especificação das buscas dos usuários de um sistema de recuperação de informação. Restringe-se, portanto, aos dois primeiros usos listados acima.

3.3 Classificação dos sistemas baseados em ontologia

Leite (2009, p.40) utiliza os termos "estrutura conceitual" e "estrutura de conhecimento" como sinônimos que referenciam genericamente objetos terminológicos auxiliares no processo de recuperação de informação, tais como tesauros e ontologias. Tendo em vista a delimitação deste trabalho, serão considerados apenas os sistemas que utilizam especificamente ontologias na sua operacionalização.

Os sistemas de recuperação de informação podem ser classificados em três categorias básicas a partir dos seguintes critérios: (1) quantidade de ontologias utilizadas; (2) fase do processo de recuperação em que a ontologia é utilizada; (3) formas de avaliação dos sistemas baseados em ontologias.

A **quantidade de ontologias** refere-se ao número de ontologias utilizadas no sistema de recuperação de informação. A maioria dos sistemas utiliza uma única ontologia, geralmente como ferramenta de padronização terminológica dos documentos e das buscas dos usuários. Porém, o *corpus* de um sistema pode conter documentos relacionados a diversos assuntos ou domínios de conhecimento, sendo necessário comportar mais de uma ontologia.

O segundo critério de classificação refere-se à **fase do processo de recuperação de informação** em que a ontologia é utilizada. Tal critério tem por objetivo identificar possibilidades de uso das ontologias nas diferentes etapas do processo de recuperação de informação. São identificadas três fases nas quais as ontologias podem ser utilizadas: *indexação*, *especificação da consulta* e *apresentação dos resultados*.

- Na fase de *indexação*, as ontologias são utilizadas geralmente para compor ou complementar os termos de indexação dos documentos. O sistema OWLIR (FININ *et al*, 2005) utiliza uma ontologia para inserir anotações semânticas nos documentos para auxiliar posteriormente no processo de indexação.
- A *especificação da consulta* pode ser realizada em uma interface que permita a seleção de termos por meio da navegação direta na ontologia, ou pela especificação de um conjunto de termos relacionados a conceitos da ontologia. O sistema OnAir (PAZ-TRILLO; WASSERMANN; BRAGA, 2005) permite

especificar as consultas em linguagem natural (texto livre), de onde são extraídos termos com o auxílio de uma ontologia. Uma ontologia pode ainda ser utilizada expandir consultas, agregando novos termos à consulta inicial do usuário a partir de seus relacionamentos.

- A utilização de ontologias permite diversificar as formas de *apresentação dos resultados*, indo além as listas ordenadas de documentos. Em alguns sistemas os resultados são apresentados como agrupamentos de documentos que se distribuem entre os conceitos da uma ontologia.

Leite (2009, p.46) identifica duas **formas de avaliação dos sistemas**: por meio da utilização das medidas de precisão e revocação (cobertura), ou pela validação junto aos usuários. Os sistemas que utilizam ontologia para marcações semânticas dos documentos ou para a expansão de consultas são geralmente avaliados utilizando as medidas de precisão e revocação. Sistemas que utilizam ontologias para a seleção de termos de busca ou na apresentação dos resultados utilizam geralmente a avaliação pela validação junto aos usuários.

3.4 Resumo e Discussão

Este capítulo apresentou o surgimento da Recuperação de Informação como uma área de pesquisa multidisciplinar, mas de particular interesse para a Ciência da Informação e a Ciência da Computação. Foram apresentados também os chamados “modelos clássicos” de recuperação de informação.

Uma característica dos modelos clássicos é considerar o processo de recuperação de informação como um sistema fechado, endógeno, no qual a semântica é resolvida no interior do próprio sistema. Já a partir da década de 1970 surgem os primeiros trabalhos que propunham a utilização de algum tipo de controle terminológico a fim de alcançarem melhores resultados. Atualmente, a utilização de ontologias na recuperação firma-se como um novo campo de pesquisa, com diversos trabalhos que abordam uma grande diversidade de propostas.

As atuais pesquisas apontam a principal utilização das ontologias como elementos auxiliares na construção das representações tanto dos documentos, vinculadas à operação de indexação, como nas buscas, por meio da expansão de consulta. Algumas poucas pesquisas utilizam as ontologias para outras finalidades, como auxiliar na inclusão de marcações

semânticas em documentos textuais ou servir de interface para a especificação de busca e de apresentação de resultados dos resultados.

Ontologias ainda é um tema de pesquisa recente tanto para a Ciência da Computação como para a Ciência da Informação, o que permite prever que muito desenvolvimento e muitas aplicações nessa área ainda estão por vir. A utilização de ontologias na recuperação de informação é um tema ainda incipiente, cujas pesquisas se iniciaram apenas no final da década de 1990. Comparando-se com toda a história de pesquisa em recuperação de informação podemos esperar ainda grandes avanços para um futuro próximo.

Referências

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 2ª ed. Addison-Wesley, 2011.

BUCKLAND, M.K. Information as thing. **Journal of the American Society of Information Science**, v.42, n.5, 1991. p.351-360.

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

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.

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

GORDON, A.S.; DOMESHEK, E.A.; Deja Vu: a knowledge-rich interface for retrieval in digital libraries. In: **IUI '98: Proceedings of the 3rd international conference on Intelligent user interfaces**. New York, ACM Press, 1998.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, 1951, p.20-32.

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.

PAYNTER, G.W.; WITTEN, I. H. A combined phrase and thesaurus browser for large document collections. In: **ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries**. London: Springer-Verlag, 2001.

SALTON,G. **The SMART Retrieval System; experiments in automatic document processing**. Englewood Cliffs, NJ: Prentice Hall, 1971.

SALTON, G.; WONG, A.; YANG, C.S. A Vector Space Model for Automatic Indexing. **Communications of the ACM**, v.18, n.11, 1975

SARACEVIC, T. Information Science. **Journal of the American Society for Information Science**, v. 50, n. 12, 1999, p.1051-1063.

STUCKENSCHMIDT,H.; van HARMELEN, F.; WAARD, A.; SCERRI, T.; BHOGAL, R.; van BUEL, J; CROWLESMITH; I.; FLUIT, C.; KAMPMAN, A; BROEKSTRA, J.; van MULLIGEN, E. Exploring large document repositories with RDF technology: The DOPE project. **IEEE Intelligent Systems**, v.19, n.3, 2004.

4
Indexação Automática baseada em Ontologia

Neste capítulo será definido o processo de indexação e as vantagens de desvantagem da automação desse processo. Utilizando a classificação dada por Lancaster, serão discutidos os diferentes tipos de indexação automática. Por fim, serão apresentadas as formas de utilização de ontologias no processo de indexação automática.

A indexação de um documento visa representar o seu conteúdo temático por meio de um conjunto de termos com o objetivo de sintetizar o seu conteúdo, ressaltando o que lhe é essencial. Um termo de indexação é constituído de uma ou mais palavras cujo significado remete preferencialmente 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.

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 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 até um tesauro ou, como veremos, uma ontologia. Após a leitura do texto, o indexador escolhe os termos mais adequados para representar o conteúdo informacional do documento.

Embora a prática da indexação possa ser regulada por políticas e princípios institucionais, o processo de indexação manual é dependente de critérios subjetivos e pessoais relacionados à formação e 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 publicados e disponibilizados, que caracterizou a "explosão informacional", justificaram estudos que buscavam soluções alternativas para auxiliar o indexador no exercício de sua atividade. As primeiras pesquisas em indexação automática aconteceram no final dos anos de 1950, época de rápido desenvolvimento das tecnologias de computação. A popularização da microinformática a partir dos anos de 1980, mas principalmente o surgimento 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.

Anderson e Perez-Carballo (2001) citam o baixo custo da indexação automática e sua facilidade de aplicação a grandes conjuntos de documentos um importante fator de incentivo ao 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 realizada por seres humanos 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.

De forma semelhante à sua classificação da indexação manual, Lancaster (2004, p.285) identifica dois tipos de indexação automática: *indexação por extração automática* e *indexação por atribuição automática*. A indexação por extração automática é realizada geralmente por meio de cálculos matemáticos de frequência das palavras encontradas no texto de um documento. Na indexação por atribuição automática é utilizado um elemento externo ao texto, com o objetivo de normalizar os termos de indexação atribuídos aos documentos.

4.1 Indexação por extração automática

A maioria dos métodos de indexação automática busca selecionar termos dos próprios textos dos documentos. Tais métodos pressupõem que os significantes, as palavras, são os únicos elementos passíveis de serem operados computacionalmente em um texto. Assim, os termos de indexação são resultantes de cálculos estatísticos e sucessivas operações algorítmicas aplicadas às palavras de um ou de um conjunto de textos.

Na literatura, é recorrente a referência a George Kingsley Zipf (1902-1950) como pioneiro nos estudos estatísticos do texto. Linguista da Universidade de Harvard, Zipf apresentou a sua lei empírica na obra *Human Behaviour and the Principle of Least Effort* (ZIPF, 1949). Analisando a obra Ulisses, de James Joyce, Zipf observou que, em um texto suficientemente longo, se listarmos as palavras em ordem decrescente de frequência, a posição de cada palavra multiplicada por sua frequência resulta um valor praticamente constante.

Utilizando a lei de Zipf como ponto de partida, Hans Peter Luhn (1896-1964), cientista da computação da IBM, sugeriu que certas palavras poderiam ser automaticamente extraídas de um texto a fim de representar o próprio texto. Porém, segundo o autor, nem todas as palavras seriam bons termos de indexação e nem todos os termos de indexação contribuem igualitariamente na representação do conteúdo informacional do texto. Luhn (SCHULTZ, 1968) propôs técnicas para identificar e atribuir pesos aos termos de indexação. Para Luhn, palavras mais significativas dos documentos são as palavras de frequência média. As palavras com frequência muito baixa seriam pouco significativas na representação do documento e as muito frequentes teriam baixo poder de representação do conteúdo informacional do documento.

Durante a década de 1960, Gerard Salton propôs o Modelo Vetorial (Seção 3.1.2) e deu início ao desenvolvimento do Sistema SMART, com o qual importantes conceitos e técnicas relacionados à recuperação de informação foram testados e avaliados. No Modelo Vetorial um documento é representado por um vetor numérico onde cada elemento representa o peso do respectivo termo de indexação. Antes de se atribuir pesos aos temos é necessário definir quais serão esses termos. Desde a sua concepção, o sistema SMART já incorporava algumas ferramentas para a extração automática de termos de indexação a partir de um *corpus*

documental. O processo de indexação do sistema SMART é dividido nas seguintes etapas (SALTON; McGILL, 1983, p.131):

1. Identificar e isolar cada palavra do texto do documento;
2. Eliminar palavras com grande frequência e pouco valor semântico (*stop words*) tais como preposições, artigos, etc.;
3. Remover afixos (prefixos e sufixos) das palavras restantes, reduzindo-as ao seu radical (processo conhecido como *stemming*);
4. Incorporar os radicais (termos) ao vetor do documento e atribuir-lhes um peso, calculado por meio da medida *tf* x *idf*.

A medida *tf* x *idf* é a forma mais conhecida para calcular os pesos dos termos de indexação. Define-se inicialmente *tf* (*term frequency*) como sendo o número de vezes que um determinado termo aparece no texto de um documento. Parte-se, portanto, do pressuposto que, retirando-se as *stop words*, a importância ou peso de um termo na representação de um documentos é proporcional à sua frequência. Porém, um termo que aparece em todos os documentos de um *corpus* terá pouca utilidade, pois possui pouco poder de discriminar um determinado documento. Portanto, para um cálculo preciso do peso de um termo é necessário uma estatística global que o caracterize em relação aos outros documentos. Esta medida, denominada "*inverse document frequency*" (*idf*), mostra como um termo é distribuído pelo *corpus*. Assim, o peso de um termo ***t*** em relação a um documento ***d*** ($w_{t,d}$) é calculado pela multiplicação da medida *tf* pela medida *idf*, isto é, *tf* x *idf*. Esta medida é utilizada para atribuir peso a cada elemento do vetor que representam um documento. Os melhores termos de indexação (os que apresentarão maior peso) são aqueles que ocorrem com uma grande frequência em poucos documentos.

A medida *tf* x *idf* é a mais referenciada e uma das mais utilizadas para atribuir peso aos termos de indexação e está diretamente relacionada ao Modelo Vetorial.

A indexação por extração automática considera que as palavras do texto de um documento comportam semântica suficiente para representar os assuntos tratados por ele. A importância ou a relevância de um determinado termo de indexação em representar o conteúdo informacional de um documento é indicada por um valor numérico, geralmente calculado pela frequência com que o termo ocorre no documento e nos demais documentos do *corpus*.

4.2 Indexação por atribuição automática

A extração de termos de um texto é uma tarefa realizada de forma relativamente satisfatória por computadores, e apresenta como vantagem a padronização e a coerência (homogeneidade), característicos dos processos algorítmicos. Porém, segundo Lancaster (2004, p.289), a maior parte da indexação realizada por seres humanos é a indexação por atribuição, utilizando um vocabulário controlado como ferramenta normalizadora.

Ainda segundo Lancaster (2004, p.19), "um vocabulário controlado é essencialmente uma lista de termos autorizados". Porém, a estrutura terminológica de um vocabulário controlado pode ir muito além de uma mera lista, podendo incluir uma "forma de estrutura semântica" destinada especialmente a: (1) controlar sinônimos optando-se por uma única forma padronizada, com remissivas de todas as outras formas; (2) diferenciar homógrafos; (3) reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si.

Lancaster (2004, p.20) classifica os vocabulários controlados em três tipos: esquemas de classificação bibliográfica (tal como a *Classificação Decimal de Dewey*), listas de cabeçalhos de assuntos e tesauros. Esses tipos apresentam os termos de forma alfabética e "sistêmica". Nos esquemas de classificação o arranjo alfabético é secundário, sobressaindo o esquema de codificação numérica dos termos em uma estrutura hierárquica. No tesauro o arranjo é alfabético, mas existe uma estrutura hierárquica implícita. Uma lista de cabeçalhos de assuntos é similar ao tesauro quanto à sua base alfabética, mas incorpora uma estrutura hierárquica "imperfeita" e não distingue claramente as relações hierárquicas das associativas. Os três tipos de vocabulário controlam sinônimos, distinguem homógrafos e agrupam termos afins, mas empregam métodos um tanto diferentes para alcançar estes objetivos.

Lancaster (2004, p.289) aponta que uma maneira "obvia" de automatizar a indexação por atribuição é criar para cada termo do vocabulário controlado um "perfil" de palavras ou expressões que costumam ocorrer nos documentos aos quais um indexador humano atribuiria esse termo. Assim, a indexação se dá em duas fazes: em uma primeira etapa extraem-se palavras ou expressões do texto por meio de técnicas estatísticas. Em uma segunda fase, partindo desse conjunto de palavras/expressões, seleciona-se no vocabulário controlado o termo cujo perfil possui certo nível de coincidente.

Uma das vantagens subjacentes ao uso de vocabulários controlados é que estes podem ser disponibilizados para os usuários de um sistema de informação, permitindo que

tenham acesso à terminologia empregada na indexação dos documentos. Isto possibilita compatibilizar a linguagem dos usuários à linguagem utilizada na representação documentos, resultando em uma recuperação mais eficiente. Além disso, a padronização terminológica proporcionada pelo uso de um vocabulário controlado permite o agrupamento lógico dos documentos de um *corpus,* criando-se grupos (*clusters*) ou classes de documentos que são representados por determinados termos de indexação.

Embora uma ontologia tenha originalmente uma finalidade diversa de uma linguagem documentária, é possível utiliza-la como tal, pois comporta um vocabulário de domínio no qual os conceitos estão explicitamente relacionados, permitindo realizar inferências e derivações semânticas.

4.3 Indexação automática baseada em Ontologia

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.

O início das pesquisas em indexação automática data da década de 1950, com os trabalhos de Hans Peter Luhn (SCHULTZ, 1968). De forma geral, os primeiros trabalhos nesse campo consideravam o texto de um documento como um elemento autônomo, cuja semântica se resolveria no interior do próprio texto. 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 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.

Embora no contexto da indexação automática existam variações nas formas de utilização de ontologias, a abordagem mais comum da *indexação automática baseada em ontologia* insere-se na categoria de *indexação por atribuição automática*, como apresentado na seção anterior. Esta abordagem está sendo utilizada no desenvolvimento do sistema OntoSmart, descrito no Capitulo 6.

4.4 Resumo e Discussão

Utilizando a classificação dos processos de indexação proposta por Lancaster, este capítulo apresentou as principais abordagens da indexação automática: indexação por extração e por atribuição. Na indexação por extração automática os termos de indexação de um documento são selecionados por meio de cálculos matemáticos realizados nas palavras do texto do próprio documento. A indexação por atribuição automática envolve um elemento externo ao documento, um vocabulário controlado, que permite uma normalização dos termos de indexação utilizados para indexar um documento ou todo um *corpus*.

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

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, n.2, 2001.

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

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

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

ZIPF, G.K. **Human Behavior and the Principle of Least Effort**. Cambridge, MA: Addison-Wesley, 1949.

5
Expansão de Consulta baseada em Ontologia

O capítulo anterior tratou do processo de indexação, responsável pela representação dos itens de informação (documentos) de um sistema de recuperação de informação. Este capítulo irá tratar da representação da necessidade de informação dos usuários, especificamente do processo denominado "expansão de consulta". Serão apresentadas as diversas formas de expansão de consultas e os métodos de utilização de ontologias como provedoras de termos de expansão.

Um sistema de recuperação de informação é um elemento mediador entre os usuários e um determinado acervo documental. O usuário interage com o sistema a fim de comunicar a sua necessidade de informação e obter documentos que possam satisfazer tal necessidade. Na maioria dos sistemas essa comunicação é feita por meio da especificação de termos que representam a necessidade do usuário. Nesse processo comunicativo entre usuário e sistema é fundamental a escolha criteriosa dos termos de busca para se recuperar documentos relevantes e ao mesmo tempo evitar itens não relevantes. Porém, sem um conhecimento de como foram representados (indexados) os documentos do *corpus* é difícil ao usuário predizer os termos que resultem em um conjunto de documentos que efetivamente atenderão à sua necessidade.

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

Spink *et al* (2001) realizaram estudos envolvendo mais de um milhão de consultas utilizando a ferramentas de busca Excite em um único dia: 16 de setembro de 1997. Constatou-se que o número médio de termos utilizados em uma consulta varia entre 2 e 3. Além disso, mais da metade dos usuários reformulam suas buscas pelo menos uma vez. Esta constatação torna clara que as consultas iniciais muitas vezes não resultam em um conjunto de documentos satisfatórios para as necessidades de informação dos usuários.

A importância e as dificuldades do processo de especificação de buscas fez surgir na área de Recuperação de Informação (*Information Retrieval*) 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. O conceito de expansão de consulta está relacionado ao conceito mais genérico de reformulação de consulta, que pode envolver também a exclusão de termos de uma consulta inicial.

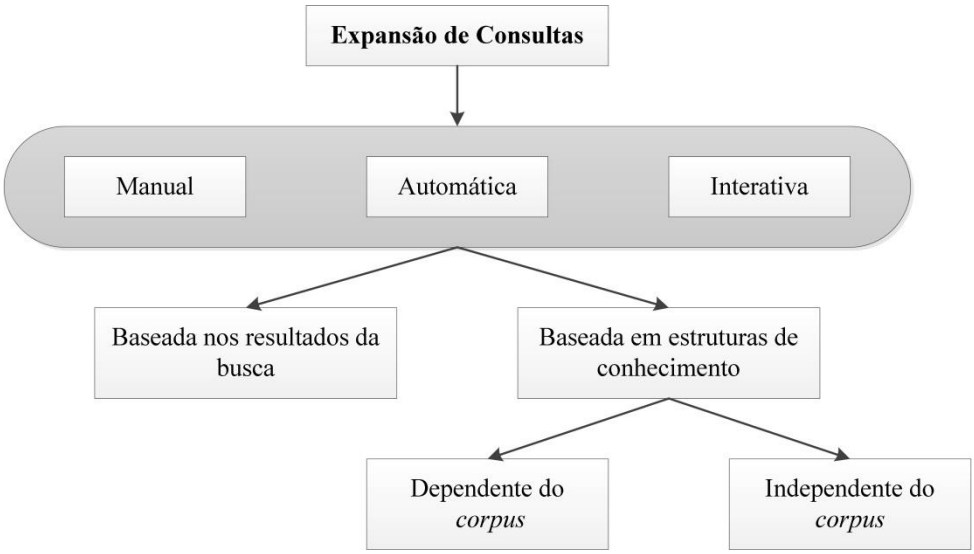
O funcionamento de um mecanismo de expansão de consulta é dependente do modelo utilizado pelo sistema de recuperação de informação. No Modelo Booleano, por exemplo, os termos de expansão são combinados com os termos da consulta original por meio de operadores booleanos. O operador OR pode ser utilizado para realizar buscas mais genéricas, com um potencial aumento na revocação (*recall*). O operador AND restringe o resultado da consulta inicial, permitindo uma maior precisão, com uma consequente redução da revocação. Nas abordagens baseadas no Modelo Vetorial, termos de expansão são adicionados à consulta original juntamente com seus respectivos pesos (ROCCHIO, 1971).

Muitas vezes os mecanismos de expansão de consultas podem ser aplicados para auxiliar o usuário na formulação da sua consulta inicial e adicionalmente ou alternativamente fazer uso de tais técnicas em etapas subsequentes, reformulando as consultas até que sejam satisfeitas suas necessidades de informação.

Efthimiadis (1996) distingue três modos diferentes de expansão de consulta, como representado na Figura 5.1. Uma reformulação é considerada *manual* (ou intelectual)sempre que o próprio usuário altera a sua consulta inicial por meio da adição de novos termos. A expansão é considerada *automática* quando o sistema gera os termos de expansão e os

adicionados à consulta original. Segundo Efthimiadis, para ser considerado automático o processo de expansão de consulta não pode ser influenciado pelo usuário nem tampouco ele pode estar ciente de sua existência. No modo *interativo* o usuário tem influência sobre a seleção de termos de expansão a partir de um conjunto de termos apresentados pelo sistema.

Figura 5.1 – Métodos de expansão de consulta



Fonte: Adaptado de Efthimiadis (1996)

Métodos de expansão de consulta podem variar ainda na forma como são gerados os termos da expansão. Como mostrado na Figura 5.1, estes termos podem ser originados dos *resultados de busca* ou de *estruturas de conhecimento*. Os métodos baseados nos resultados da busca selecionam os termos de expansão a partir dos documentos resultantes da consulta inicial. Nesse caso, a eficácia da expansão da consulta depende fortemente da qualidade da consulta original. Essa dependência não existe nos modelos de expansão baseados em *estruturas de conhecimento*.

As estruturas de conhecimento podem ser *dependentes do corpus* ou *independentes do corpus*. Mecanismos dependentes do *corpus* analisam os documentos do acervo documental a fim de selecionar os termos que serão utilizados para a expansão da consulta. Métodos independentes do *corpus* contam com estruturas de conhecimento que não apresentam relação com os documentos. São exemplos dessas estruturas: léxicos, glossários, dicionários, tesauros, ontologias.

5.1 Expansão de consultas baseada nos resultados da busca

A expansão de consultas baseadas nos resultados da busca está relacionado ao processo denominado *Relevance Feedback*. Este processo parte da ideia de que embora seja difícil formular uma primeira consulta eficiente, é fácil julgar a relevância dos documentos recuperados.

Sistemas de recuperação de imagens são bons exemplos da eficiência do mecanismo de *Relevance Feedback*. Neste domínio, as dificuldades do usuário em formular sua primeira consulta são maiores, em face da complexidade em traduzir em palavras as características e propriedades das imagens de interesse. Mas, por outro lado, o usuário tem condições de rapidamente julgar a relevância das imagens apresentadas nos resultados, iniciando assim um processo de refinamento da busca. O processo de *Relevance Feedback* pode ser resumido nos seguinte passos (MANNING, RAGHAVAN; SCHÜTZE, 2008, p.178):

- O usuário formula uma consulta e submete ao sistema;
- O sistema retorna um conjunto inicial de documentos;
- O usuário marca como relevante ou não-relevante alguns dos documentos recuperados e submete novamente ao sistema;
- O sistema calcula uma melhor representação da necessidade de informação baseada no *feedback* do usuário.
- O sistema apresenta um novo conjunto de documentos presumivelmente apresentado um aumento da precisão dos resultados.

Essa interação com o sistema pode se repetir até que o usuário esteja satisfeito como o conjunto de documentos resultantes.

Ruthven e Lalmas (2003) apresentam uma estudo aprofundado dos mecanismos de *Relevance Feedback.* Basicamente, existem dois tipos: *User Relevance Feedback* e *Pseudo Relevance Feedback*. No *User Relevance Feedback*, o usuário pode indicar (marcar) os documentos resultantes de uma consulta como relevantes ou não-relevantes e submeter essa nova informação ao sistema, que a utiliza na modificação da consulta original, adicionando novos termos e/ou alterando os pesos dos termos da consulta inicial a fim melhorar a eficácia da consulta. O *Pseudo Relevance Feedback Feedback* não confia na informação de relevância fornecida pelo usuário e utiliza os documentos mais bem ranqueados na lista de resultados

para aperfeiçoar a consulta. Esta técnica depende fortemente da qualidade da consulta inicial e de sua aptidão em recuperar documentos relevantes.

Embora os mecanismos de *Relevance Feedback* provem ser eficazes para melhorar resultados da recuperação, estes sofrem de diversos inconvenientes. Eles são eficazes somente se os documentos na coleção contiverem uma determinada quantidade de texto dos quais os termos da expansão possam ser obtidos. Este não é o caso em bases de dados de referências, por exemplo, onde os metadados dos documentos geralmente compreendem somente uma quantidade pequena de texto livre. Além disso, o método de *Relevance Feedback* é somente aplicável se a consulta original do usuário resultar em um conjunto com um número razoável de documentos. Métodos de *Relevance Feedback* não podem ser aplicados na formulação da consulta inicial, pois nenhuma resultado estará disponível.

Os mecanismos de *User Relevance Feedback* são dependentes da voluntariedade dos usuários em fornecer o seu parecer sobre a relevância dos documentos recuperados. Segundo Spink *et al* (2000), na maioria das vezes os usuários são relutantes em fazer isso. Os autores analisaram o comportamento de usuários de mecanismos de busca Web baseado em dados do arquivo de *log* do buscador Excite. Essa ferramenta de busca possuía o recurso "*More Like This*" como uma opção de *Relevance Feedback*. Embora reconhecidamente útil, o mecanismo de *Relevance Feedback* raramente era utilizada pelos usuários.

5.2 Expansão de consultas baseada em estruturas de conhecimento dependentes do *corpus*

Em vez de se obter termos de expansão a partir dos documentos resultantes de uma busca, os métodos dependentes do *corpus* utilizam a estrutura de todos os documentos para identificar tais termos. Para este propósito, dependências estatísticas entre termos são calculadas por meio da aplicação de cálculos de co-ocorrência. Uma forma simples de utilizar dados de co-ocorrência é identificar nos documentos termos de indexação que se assemelham aos termos de uma determinada consulta com o objetivo utilizá-los como termos de expansão. Dados de co-ocorrência são utilizados também em métodos de agrupamento (*clustering*) (LIU; NATARAJAN; CHEN, 2011), na geração de matrizes de co-ocorrência, ou ainda na construção automática de tesauros de similaridade (*similarity thesaurus*) (QIU, 1995).

Jing e Croft (1994) usaram os termos de um tesauro criado automaticamente para a expansão de consulta. Os autores relataram uma melhoria significativa na eficácia da recuperação. Resultados semelhantes foram obtidos por Qiu e Frei (1993), que reforçam a eficácia de um tesauro de similaridade para expansão de consulta, conduzindo a uma melhoria nos índices de revocação e precisão.

Comparado com os mecanismos de *Relevance Feedback*, métodos de expansão baseados em co-ocorrencia possuem a vantagem de poderem ser aplicados na formulação da consulta inicial. Além disso, as estruturas de conhecimento podem ser criadas automaticamente e são facilmente adaptáveis a um *corpus* dinâmico (KRISTENSEN, 1993).

5.3 Expansão de consultas baseada em estruturas de conhecimento independentes do *corpus*

Estruturas de conhecimento constituem fontes especialmente promissoras para a geração dos termos de expansão nos casos em que mecanismos de *Relevance Feedback* e de co-ocorrência não são aplicáveis. Bhogal *et al* (2007) salientam que as estruturas de conhecimento independentes do *corpus* são especialmente úteis se o número de documentos for pequeno ou se os seus documentos contiverem pouco texto livre. Neste caso, os mecanismos de *Relevance Feedback* e os mecanismos baseados em co-ocorrência provavelmente não serão muito eficazes. A aplicabilidade dos métodos de expansão de consulta baseados em estruturas de conhecimento, por outro lado, independem do tamanho do *corpus*.

Outra vantagem do uso de estruturas de conhecimento para a expansão da consulta é sua disponibilidade a qualquer momento no processo de busca. Ao contrário dos mecanismos baseados em *Relevance Feedback*, a consulta inicial pode já se beneficiar deste tipo de expansão, pois os termos não são derivados de resultados da busca inicial. No entanto, o desenvolvimento de estruturas de conhecimento adequadas para fins de expansão de consulta pode ser um processo de alto custo. Como afirmado por Harman (1988) e Greenberg (2001), o desenvolvimento de mecanismos de expansão de consulta independentes do *corpus* muitas vezes é dificultada pela disponibilidade limitada de tesauros ou ontologias. No entanto, com o surgimento e o desenvolvimento da Web Semântica, grande número de ontologias está

atualmente em desenvolvimento ou já estão disponíveis na Web, o que pode resultar um impulso significativo a esse tipo expansão de consultas.

5.4 Expansão de consultas baseada em ontologia

Uma ontologia, considerada uma estrutura de conhecimento independente do *corpus*, pode ser utilizada na expansão das consultas inicialmente formuladas pelos usuários por meio da inserção de novos termos derivados dos relacionamentos entre os seus conceitos.

A partir de uma interface adequada, as ontologias podem servir também como ferramentas para a seleção dos termos que irão compor a consulta inicial do usuário. Isso permite a uma pessoa leiga em um determinado domínio ou assunto consiga realizar consultas pertinentes em um sistema de recuperação de informação, ao mesmo tempo em que se familiariza com a terminologia do domínio de interesse. Katifori *et al* (2007) apresentam um estudo aprofundado e abrangente sobre os métodos de visualização de ontologias.

Dey *et al* (2005) usaram ontologias de domínio para a implementação de mecanismos de expansão de consulta. Para a determinação das condições de expansão, foi calculada a distância semântica entre os termos de consulta e os conceitos de duas ontologias: uma ontologia sobre vinhos e outra sobre plantas. Como resultado de suas experiências em buscas na Web utilizando o Google, os autores relatam um aumento na precisão das consultas que foram expandidas com os termos das ontologias. A precisão das consultas expandidas com a ontologia apresentou um aumento significativo da precisão dos resultados.

Sack (2005) também demonstrou como uma ontologia de domínio pode aumentar a eficiência de um sistema de recuperação de informação tradicional. Sua pesquisa se apoiou em uma base de dados bibliográfica e uma ontologia do domínio de problemas NP-completos. Essa ontologia foi utilizada na fase de formulação de consulta para fins de expansão e para resolução de ambiguidades. Em um modo interativo de expansão, termos semanticamente relacionados como sinônimos, termos específicos e termos genéricos eram sugeridos aos usuários. O autor aponta as vantagens do uso de uma ontologia ao fornecer aos usuários um conhecimento contextualizado.

5.5 Resumo e Discussão

Este capítulo apresentou as diversas técnicas de expansão de consulta, considerando a proveniência dos termos de expansão. A expansão de consulta é definida como manual quando o próprio usuário refaz sua consulta inicial acrescentando novos termos. A expansão de uma consulta pode ser feita partindo-se do julgamento de relevância (ou não-relevância) do usuário, após ter obtido um primeiro conjunto de documentos resultantes de uma consulta inicial (*Relevance Feedback*). Os termos de expansão podem também ser derivados de estruturas construídas a partir dos resultados de uma consulta ou considerando todo o *corpus* documental de um sistema de recuperação de informação.

Partindo-se do pressuposto que uma ontologia comporta a terminologia de uma determinada área do conhecimento, é natural considerar sua eficiência como ferramenta para especificação ou expansão de consultas. Tais ferramentas seriam de grande utilidade para usuários com pouco ou nenhum conhecimento de uma determinada área ou assunto. Além disso, dependendo da interface de apresentação da ontologia, o usuário poderia conhecer rapidamente os conceitos envolvidos em um determinado domínio.

De qualquer forma, a utilização de ontologias como suporte à especificação ou expansão de consultas abre novas perspectivas e novos caminhos para a pesquisa em recuperação de informação.

Referências

BHOGAL, J., Macfarlane, A.; Smith, P. A review of ontology based query expansion. **Information Processing and Management**, v.43, n.4, 2007.

DEY, L.; SINGH, S.; RAI, R.; GUPTA, S. Ontology aided query expansion for retrieving relevant texts. In: **Proceedings 3rd International Atlantic Web Intelligence Conference**. Lodz, Poland, 2005.

EFTHIMIADIS, E. N. Query expansion. In: WILLIAMS, M.E. **Annual Review of Information Science and Technology-ARIST**. Medford, N.J.: Information Today, 1996.

GREENBERG, J. Automatic query expansion via lexical-semantic relationships. **Journal of the American Society for Information Science and Technology**, v.52, n.5, 2001.

HARMAN, D. (1988): Towards interactive query expansion. In: **Proceedings 11th annual international ACM Conference on Research and Development in Information Retrieval**, Grenoble, France, 1988.

JING, Y.; CROFT, W. B. An association thesaurus for information retrieval. In: **Proceedings RIAO 94**. New York, 1994.

KATIFORI, A; HALATSIS, C.; LEPOURAS, G.; VASSILAKIS, C.; GIANNOPOULOU, E. Ontology visualization methods - a survey. **ACM Computing Surveys**, v.39, n.4, 2007.

KRISTENSEN, J. Expanding end-users' query statements for free text searching with a search-aid thesaurus. **Information Processing and Management**, v.29, n.6, 1993.

LIU, Z.; NATARAJAN, S.; CHEN, Y. Query Expansion Based on Clustered Results. **Proceedings of the VLDB Endowment**, v.4, n.6, 2011.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. New York: Cambridge University Press, 2008.

QIU, Y. **Automatic Query Expansion Based on a Similarity Thesaurus**. 1995. Ph.D. Thesis (Doctor of Technical Sciences) - Swiss Federal Institute of Technology Zurich, ETH Zurich. 1995.

QIU, Y; FREI, H. P. Concept based query expansion. In: **Proceedings 16th annual international ACM SIGIR 2009**. Pittsburgh, US, 1993.

ROCCHIO, J. Relevance feedback in information retrieval. In: SALTON, G.: **The SMART Retrieval System**: experiments in automatic document processing. Englewood Cliffs, US, Prentice-Hall, 1971.

RUTHVEN, I.; LALMAS, M. A survey on the use of relevance feedback for information access systems. **The Knowledge Engineering Review**, n.18, v.2, 2003.

SACK, H. NPBibSearch: An ontology augmented bibliographic search. In: **Proceedings 2nd Italian Semantic Web Workshop**. Trento, Italy, 2005.

SPINK, A.; WOLFRAM, D.; JANSEN, B.J.; SARACEVIC, T. Searching the Web: The public and their queries. **Journal of the American Society for Information Science and Technology**, v.52, n.3, 2001.

SPINK, A.; JANSEN, B.J.; OZMULTU, H.C. Use of query reformulation and relevance feedback by Excite users. **Internet Research: Electronic Networking Applications and Policy**, v.10, n.4, 2000.

6
OntoSmart:
um sistema de recuperação de informação baseado em ontologia

No Capítulo 4 foram apresentadas algumas formas de utilização de ontologias no processo de indexação automática. No Capítulo 5 foi apresentado o conceito de expansão de consulta e os métodos de utilização de ontologias como provedoras de termos de expansão. Esses dois capítulos fornecem subsídios para o desenvolvimento do sistema que será descrito a seguir.

Neste capítulo será apresentada a base teórica utilizada no desenvolvimento de um sistema de recuperação de informação baseado em ontologia denominado OntoSmart. Incialmente, o Modelo Espaço Vetorial, já visto resumidamento na Seção 3.1.2, é abordado de forma mais aprofundada, pois constitui a estrutura básica na qual o sistema se assenta. Alguns conceitos matemáticos serão introduzidos a fim prover um suporte adequado à operacionaliação do sistema.

O sistema OntoSmart utiliza a estrutura formal do Modelo Espaço Vetorial, associado ao uso de ontologias como ferramenta de normalização da terminologia durante o o processo de criação dos vetores representativos dos documentos e das buscas.

Embora o sistema OntoSmart já estaja em desenvolvimento e com alguns resultados disponíveis, optou-se por apresentar apenas os conceitos e a metodologia utilizada na sua implementação. Por causa da limitação de tempo para a realização desse trabalho, não foi possível realizar testes comparativos necessários para aferir o desempenho do sistema. Tais testes terão que ficar para um futuro próximo.

6.1 Conceitos básicos

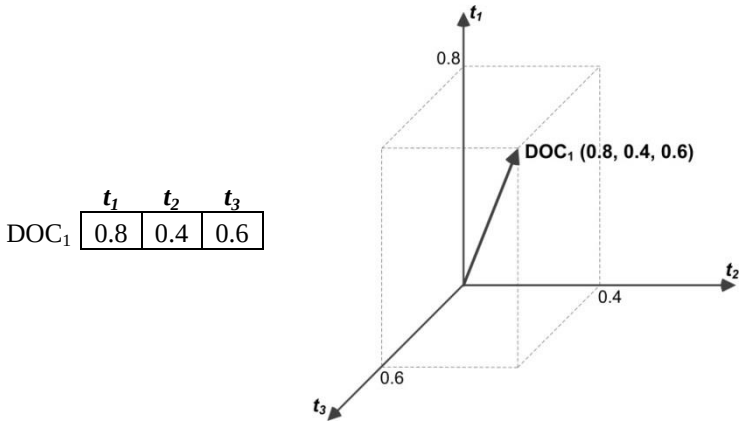
Como mencionado anteriormente, o Modelo Espaço Vetorial forma a alicerce no qual se assenta o sistema OntoSmart. As ontologias são utilizadas como auxiliares no processo de composição dos vetores que representam os documentos e as buscas dos usuários.

6.1.1 Modelo Espaço Vetorial

O Modelo Espaço Vetorial já foi apresentado resumidamente no Capítulo 3. Nesta seção, este modelo será abordado mais detalhadamente, pois é usado como base para o desenvolvimento do sistema OntoSmart.

No Modelo Vetorial 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). Os pesos mais próximos de um (1) indicam termos mais relevantes na representação do conteúdo informacional do documento. Para possibilitar sua visualização em um espaço cartesiano tridimensional, é apresentado na Figura 6.1, a título de exemplo, um documento **DOC_1** representado por três termos de indexação (***t_1***, ***t_2*** e ***t_3***) de pesos 0.8, 0.4 e 0.6, respectivamente.

Figura 6.1 – Representação vetorial de um documento com três termos de indexação



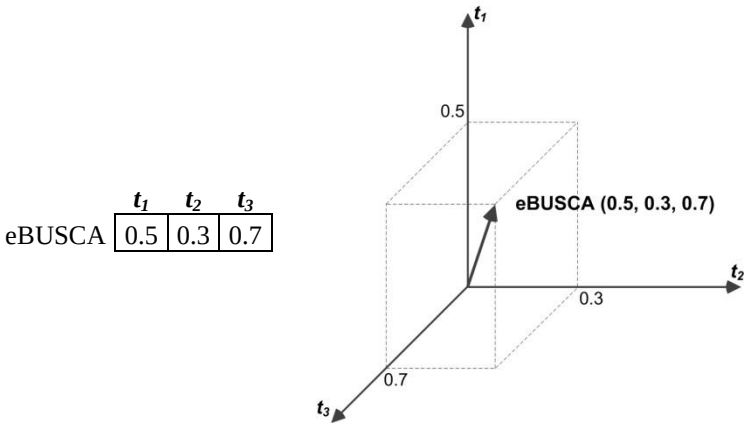
	t_1	t_2	t_3
DOC_1	0.8	0.4	0.6

Fonte: desenvolvida pelo autor

Da mesma forma que os documentos, uma expressão de busca 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 6.2 ilustra a

representação vetorial uma expressão de busca (**eBUSCA**) contendo três termos (***t₁***, ***t₂*** e ***t₃***) com pesos 0.5, 0.3 e 0.7, respectivamente.

Figura 6.2 – Representação vetorial de uma expressão de busca



	t_1	t_2	t_3
eBUSCA	0.5	0.3	0.7

Fonte: desenvolvida pelo autor

A representação gráfica (visual) de um espaço vetorial só é possível utilizando no máximo três termos. Um sistema real irá conter um grande número de documentos, cada qual representado por um conjunto de termos de indexação de certo tamanho. Um *corpus* contendo *n* documentos e *i* termos de indexação pode ser representado por uma matriz na qual cada linha representa um documento e cada coluna representa a associação de um termo específico para representação dos documentos.

	t_1	t_2	t_3	…	t_n
DOC₁	$w_{1,1}$	$W_{1,2}$	$W_{1,3}$	…	$w_{1,n}$
DOC₂	$w_{1,2}$	$w_{2,2}$	$W_{2,3}$	...	$W_{2,n}$
⋮	⋮	⋮	⋮	⋮	⋮
DOC$_m$	$W_{m,1}$	$w_{m,1}$	$w_{m,1}$	...	$w_{m,n}$

onde $w_{i,j}$ representa o peso do *i*-ésimo termo para o *j*-ésimo documento.

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 ***t*** dimensões, a similaridade (***sim***) entre um documento d_j e uma expressão de busca q é calculada utilizando a seguinte fórmula (BAEZA-YATES; RIBEIRO-NETO, 2011, p.78):

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \bullet \vec{q}}{|\overrightarrow{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

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

A fórmula de similaridade (***sim***) mede o cosseno do ângulo formado entre dois vetores, tendo como resultado um valor entre 0 e 1. Quanto menor for o ângulo, maior será a similaridade entre eles. A similaridade entre o documento **DOC_1** e a expressão **eBUSCA**, apresentados respectivamente nas Figuras Figura 6.1 e Figura 6.2, é calculada como:

	t_1	t_2	t_3
DOC_1	0.8	0.4	0.6

	t_1	t_2	t_3
$eBUSCA_1$	0.5	0.3	0.7

$$sim(\mathbf{DOC_1}, \mathbf{eBUSCA}) = \frac{(0.8 \times 0.5) + (0.4 \times 0.3) + (0.6 \times 0.7)}{\sqrt{(0.8^2 + 0.4^2 + 0.6^2)} \times \sqrt{(0.5^2 + 0.3^2 + 0.7^2)}} \cong 0.96$$

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 de recuperação de informação 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.

O Modelo Vetorial fornece um formalismo matemático bastante consistente, o que permite o desenvolvimento de sistemas poderosos e robustos. Aliada à sua relativa facilidade de implementação, este modelo de recuperação de informação foi uma escolha natural para o desenvolvimento do sistema OntoSmart.

6.1.2 Distância Semântica (*ds*)

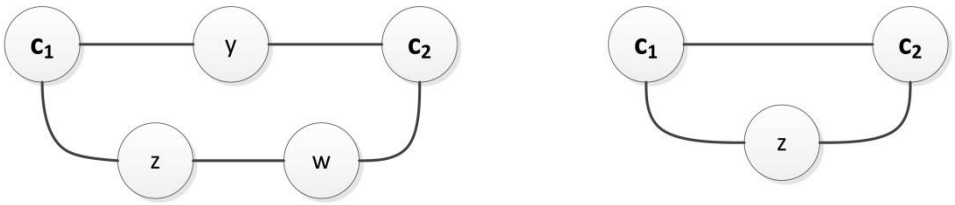
Uma ontologia O = (C, R) é composta por um conjunto de conceitos C={c_1, c_2,.., c_n} interconectados por um conjunto de relacionamentos em R={r_1, r_2, ..., r_n}. Define-se incialmente a *distância semântica* entre dois conceitos:

Definição 1

*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 6.3 (a), 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. Já na Figura 6.3 (b) *ds*(c_1, c_2)=1, pois os conceitos **c_1** e **c_2** são adjacentes, separados por um único relacionamento.

Figura 6.3 – Ilustração do conceito de *distância semântica* (*ds*)



(a) **(b)**

Fonte: desenvolvida pelo autor

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**.

6.1.3 Valor Semântico (*vs*)

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.

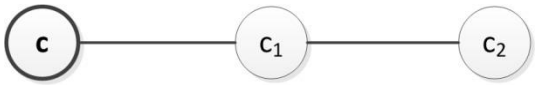
Definição 2

*Dado um conceito **c** de uma ontologia, o valor semântico (**vs**) de cada um 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**.*

Na figura abaixo temos um conceito **c** relacionado aos conceitos **c$_1$** e **c$_2$**. A distância semântica de **c$_1$** em relação a **c** é igual a 1. A distância semântica de **c$_2$** em relação à **c** é igual a 2.



***ds*(c$_1$, c)=1** **_ds_(c$_2$, c)=2**

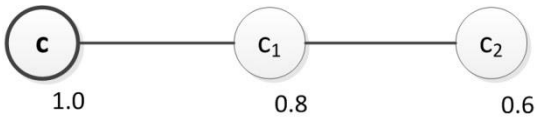
Considerando o parâmetro ***p*** = 0.2, pode-se calcular o valor semântico (***vs***) dos conceitos **c$_1$** e **c$_2$** em relação à **c** como:

***vs*(c$_1$, c) = 1 – [*ds*(c$_1$, c)×0.2] = 0.8**

***vs*(c$_2$, c) = 1 – [*ds*(c$_2$, c)×0.2] = 0.6**

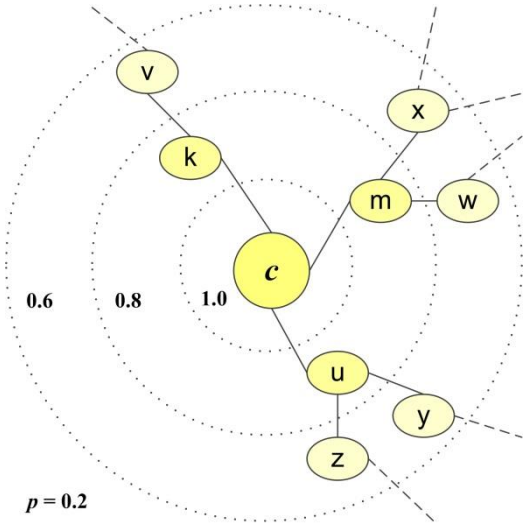
Como apresentado anteriormente, a distância semântica (***ds***) de um conceito em relação a ele próprio é igual a zero. Assim, o valor semântico (***vs***) de um conceito em relação a ele mesmo é igual a 1.

Utilizando o conceito **c** como referência (conceito central) e ***p*** = 0.2, é possível atribuir valores aos diversos conceitos de uma ontologia, como exemplificado na figura abaixo:



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 6.4 apresenta uma ilustração mais ampla da aplicação dos conceitos de distância semântica (***ds***) e valor semântico (***vs***), considerando o parâmetro ***p***=0.2.

Figura 6.4 – Ilustração do conceito de *valor semântico* (*vs*)



Fonte: desenvolvida pelo autor

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 6.4 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 6.4 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 6.4 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).

6.2 O Sistema OntoSmart

Nesta seção será descrito o funcionamento do sistema OntoSmart. Algumas de suas funcionalidades não estão formalmente especificadas, estando ainda ao nível de ideias que buscam um amadurecimento a fim de serem futuramente implementadas.

Para um melhor entendimento do funcionamento do sistema, foram utilizadas algumas ilustrações que simulam telas/janelas de entrada de dados. Os textos existentes nessas ilustrações estão no idioma inglês. Por se tratar de um sistema Web, buscou-se utilizar uma língua extensivamente conhecida e utilizada, como é a língua inglesa.

6.2.1 Cadastro de ontologia

Toda ontologia deve ser formalmente cadastrada no sistema, o que permite agregar algumas informações, tais como nome, a referência ao arquivo OWL, seu idioma e um texto livre contendo uma descrição ou algumas observações sobre a ontologia. O sistema permite o cadastro de um número ilimitado de ontologias.

Figura 6.5 – Cadastro de Ontologia

Fonte: desenvolvida pelo autor

As ontologias devem estar definidas em OWL (*Web Ontology Language*), linguagem que permite representar os aspectos semânticos e os relacionamentos existentes entre os conceitos de um domínio. É uma linguagem recomendada pelo W3C e possui ampla aceitação na comunidade acadêmica, sendo possível localizar e utilizar livremente ontologias OWL dos mais vários domínios em diversos repositórios disponíveis na Web, tais como Swoogle[4], BioPortal[5], SchemaPedia[6], TONES Ontology Repository[7].

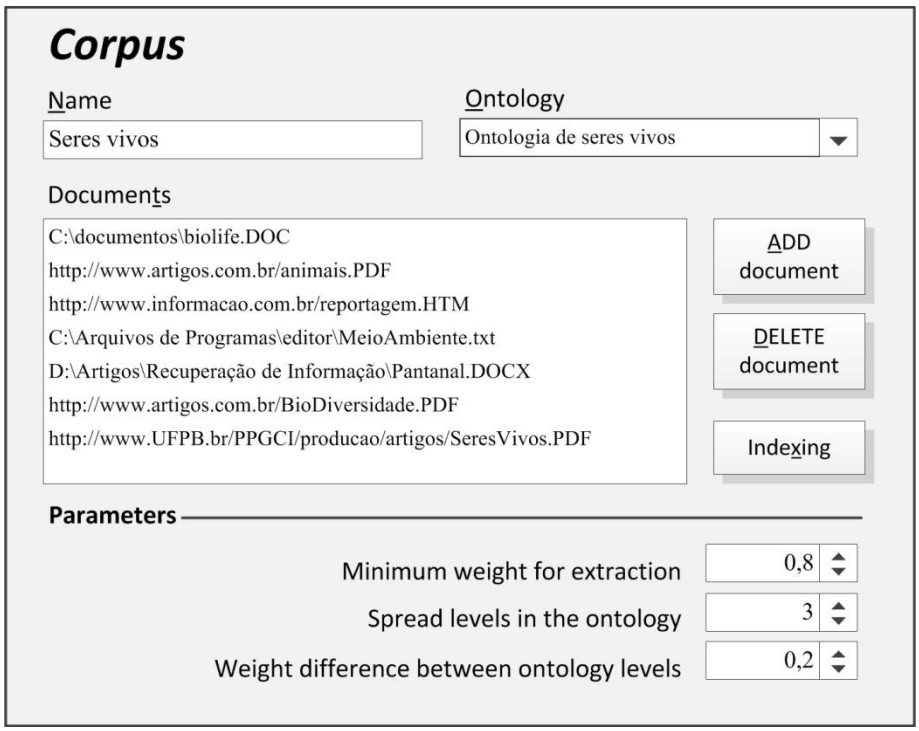
6.2.2 Definição do *corpus*

O sistema OntoSmart permite a definição de diversos *corpura*, cada qual contendo um número arbitrário de documentos. Durante o cadastro de um *corpus* deve ser definido um nome, a ontologia ao qual o *corpus* estará vinculado e o seu conjunto de documentos. Um documento é referenciado por meio do endereço de um arquivo, que pode estar localização no disco do computador local ou de uma rede, ou pode ainda ser uma página HTML referenciada por sua URL.

Todo *corpus* está vinculado a uma ontologia de domínio, definida em um determinado idioma. Portanto, os documentos de um *corpus* devem ser do mesmo idioma e do mesmo domínio da ontologia.

[4] http://swoogle.umbc.edu/
[5] http://bioportal.bioontology.org/
[6] http://schemapedia.com/
[7] http://owl.cs.manchester.ac.uk/repository/

Figura 6.6 – Cadastro de *Corpus*



Fonte: desenvolvida pelo autor

Na Figura 6.6 os parâmetros (*Parameters*) são utilizados nos processos de indexação do *corpus* e expansão da consulta. *Minimum weight for extraction* é utilizado para limitar o número de termos de indexação extraídos de um documento por meio de métodos estatísticos, como será visto na próxima seção. *Spread levels in the ontology* corresponde ao parâmetro ***k*** e w*eight difference between ontology levels* corresponde ao parâmetro ***p***, vistos na Seção 6.1.3.

6.2.3 Indexação dos documentos

A indexação tem por objetivo criar uma representação do conteúdo informacional de cada documento do *corpus*. Assim como no Modelo Vetorial (Seção 3.1.2), no sistema 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 sistema OntoSmart os pesos são calculados por meio da utilização da ontologia associada ao *corpus* do qual o 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 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 um método de indexação por extração automática. Os pesos dos termos são calculados utilizando a medida *tf* x *idf*, definida inicialmente por Salton e Yang (1973), apresentado na Seção 4.1.

Seguindo o exemplo iniciado nas seções anteriores e considerando um determinado documento, ao final do processo de extração poderia se obter, por exemplo, os seguintes termos de indexação e seus respectivos pesos:

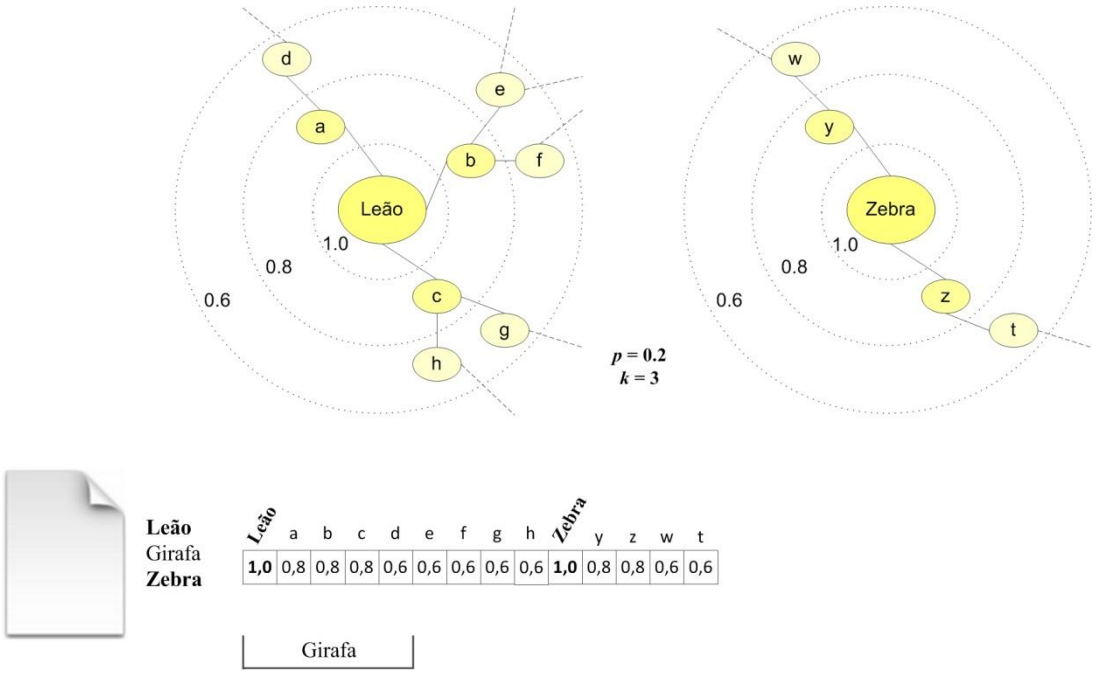
Leão	**0.9**
Girafa	**0.85**
Zebra	**0.8**
Macaco	0.5
Floresta	0.4
Savana	0.35

Nesse exemplo foram extraídos do documento seis termos com seus respectivos pesos. Considerando o parâmetro *Minimum weight for extraction*, definido no cadastro do *corpus* (Figura 6.6), serão utilizados apenas os termos cujo peso seja maior ou igual a 0.8. Portanto, o documento será representado apenas pelos três temos de maior peso, desconsiderando os termos com peso menores que 0.8.

Os termos extraídos do documento serão considerados sucessivamente como conceitos centrais da ontologia associada ao *corpus*. A ontologia terá duas funções: (1) expandir o conjunto de termos de indexação; (2) atribuir pesos a cada um dos termos.

No exemplo da Figura 6.7, verifica-se que apenas dois dos três termos extraídos do texto do documento têm relação com conceitos da ontologia. São eles: “Leão” e “Zebra”. 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 conceito, quanto maior a sua distante semântica (***ds***) do “conceito central”, menor será o seu valor semântico (***vs***).

Figura 6.7 – Representação vetorial de um documento utilizando ontologia



Fonte:desenvolvido pelo autor

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 *Weight difference between ontology levels*, que no exemplo possui valor igual a 0.2. Assim, os termos **a**, **b** e **c** receberão o valor 0.8 e os termos **d**, **e**, **f**, **g** e **h** 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.

Assim, considerando os parâmetros *Weight difference between ontology levels* (***p***) igual a 0,2 e *Spread levels in the ontology* (***k***) igual a 3, constrói-se o vetor que representa o documento. Esse processo é repetido para cada um dos documentos do *corpus*.

6.2.4 Criando um repositório de termos

No exemplo da Figura 6.7 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 atribui um peso de valor relativamente alto.

No sistema OntoSmart esses termos serão armazenados em um tipo de repositório, formando um conjunto de potenciais conceitos a serem inseridos na ontologia. Se um determinado termo for repetidamente extraído dos documentos ele poderá ser convertido em um conceito da ontologia relacionada ao *corpus*. Considerando o exemplo, se o termo "Girafa" for extraído de diferentes documentos, esse termo pode vir a ser um novo conceito da ontologia associada ao *corpus*.

Esse mecanismo de povoamento de ontologias deverá ainda ser formalizado antes de ser incorporado no sistema OntoSmart.

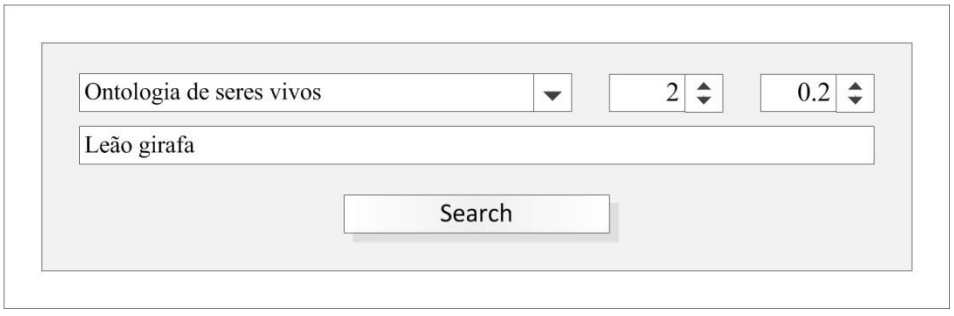
6.2.5 Especificação da busca

No sistema OntoSmart, assim como no Modelo Vetorial, 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

Na Figura 6.8 é apresentada uma figura ilustrativa de uma interface para especificação de busca.

Figura 6.8 – Especificação da busca



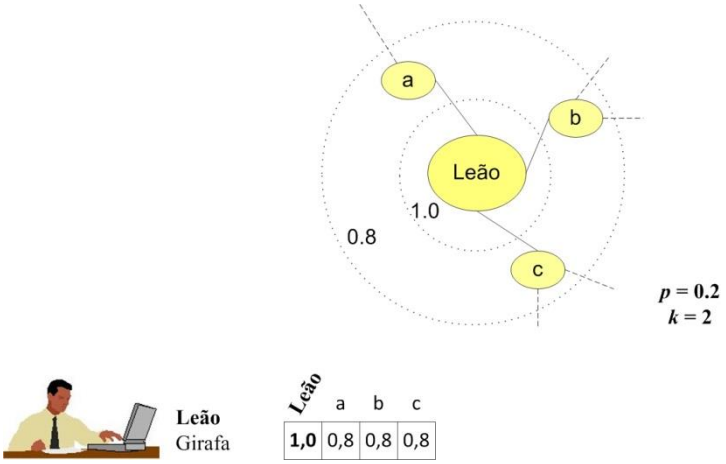
Fonte: elaborada pelo autor

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.

Como exemplo, considere uma consulta na qual o usuário utilizou dois termos: “Leão” e “Girafa”. Fazendo-se uma busca na ontologia selecionada, 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). O termo “Girafa” será descartado.

Figura 6.9 – Representação vetorial de uma expressão de busca utilizando ontologia



Fonte: elaborado pelo autor

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 6.9.

Uma possibilidade a ser desenvolvida no sistema OntoSmart é a apresentação da ontologia em uma forma gráfica e visual. Nesse caso a busca seria formulada por meio da seleção dos termos de busca diretamente na ontologia, o que evitaria erros e possibilitaria ao usuário se familiarizar com a terminologia do domínio de seu interesse.

6.2.6 Executando uma busca

A utilização de uma mesma representação tanto para os documentos como para a expressão de busca permite calcular o grau de similaridade entre a busca e cada um dos documentos do *corpus*. Como já foi visto anteriormente, em um espaço vetorial contendo ***N*** dimensões, a similaridade (***sim***) entre um documento $\boldsymbol{d_j}$ e uma expressão de busca ***q*** é calculada como definido na Seção 6.1.1.

Seguindo os exemplos ilustrativos dos processos de representação vetorial dos documentos e das busca (Figura 6.7 e Figura 6.9), a figura abaixo que apresenta um vetor de busca e um vetor de um documento.

Leão	a	b	c
1,0	0,8	0,8	0,8

Leão	a	b	c	d	e	f	g	h	Zebra	y	z	w	t
1,0	0,8	0,8	0,8	0,6	0,6	0,6	0,6	0,6	**1,0**	0,8	0,8	0,6	0,6

A partir de uma expressão de busca (consulta), representado por um vetor numérico, é calculada a similaridade para cada um dos vetores dos documentos do *corpus*. No exemplo aqui apresentado, a similaridade entre a busca e o documento é calculada como:

$$\frac{(1.0\times1.0)+(0.8\times0.8)+(0.8\times0.8)+(0.8\times0.8)}{\sqrt{1.0^2+0.8^2+0.8^2+0.8^2}\times\sqrt{1.0^2+0.8^2+0.8^2+0.8^2+0.6^2+0.6^2+0.6^2+0.6^2+0.6^2+1.0^2+0.8^2+0.8^2+0.6^2+0.6^2}}\cong 0.6507$$

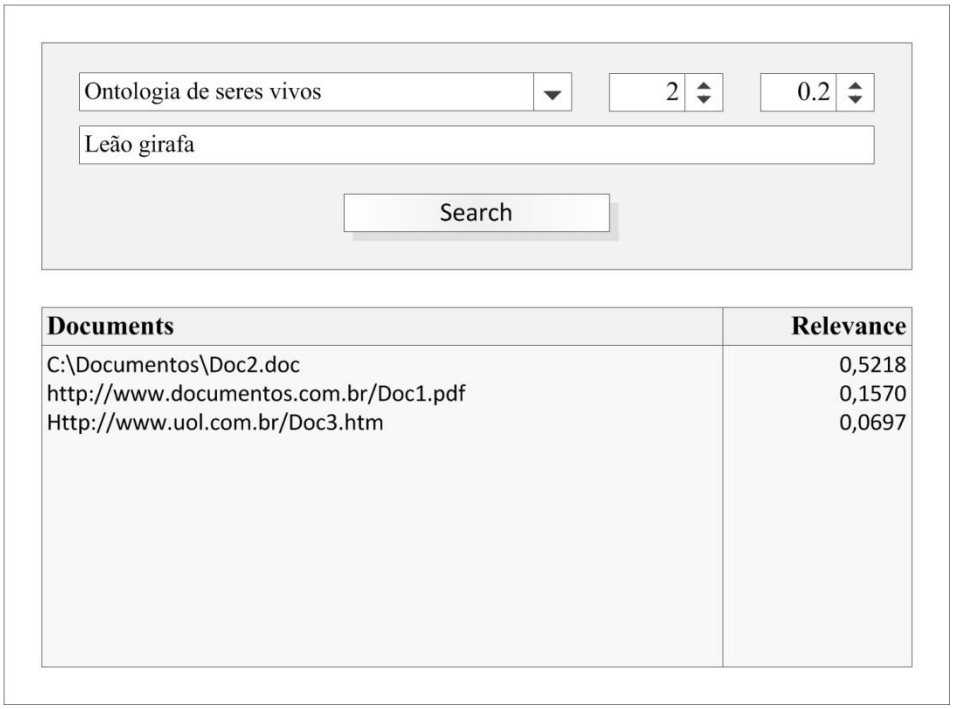
Portanto, o grau de similaridade entre a expressão de busca e o documento é aproximadamente igual a **0.6507**.

É importante observar o fato que antes de realizar uma busca o usuário do sistema precisa definir a ontologia do domínio de seu interesse. Considerando que é possível existir no sistema diversos *corpora* relacionados a uma mesma ontologia. O processo de busca se dará nos documentos de todos os *corpora* que estiverem relacionados à ontologia definida na busca do usuário.

6.2.7 Resultados de uma busca

Utilizando-se o calculo de similaridade entre uma expressão de busca e cada documento do *corpus* é possível apresentar os documentos em ordem decrescente de similaridade (relevância). A Figura 6.10 apresenta uma ilustração simplificada de uma interface de apresentação dos resultados de uma busca.

Figura 6.10 – Resultado de busca



Fonte: elaborada pelo autor

A utilização do Modelo Vetorial será possível incorporar ao sistema OntoSmart alguns parâmetros que permitam restringir os resultados de uma busca a um número máximo de documentos ou determinar um valor mínimo de relevância dos documentos resultantes.

6.3 Resumo e Discussão

Neste capítulo foi apresentado o sistema OntoSmart, que utiliza como base a estrutura formal proposta no Modelo Espaço Vetorial. Nesse sistema, os processos de criação do vetor de representação de cada documento e do vetor da busca utilizam ontologias como estrutura terminológica auxiliar na expansão de um conjunto inicial de termos. A partir de termos extraídos estatisticamente dos documentos, termos derivados de uma ontologia são inseridos no vetor numérico que representa cada documento. De maneira similar, os termos inicialmente utilizados na expressão de busca do usuário são complementados com novos termos derivados dos conceitos de uma ontologia e utilizados na criação do vetor de busca.

Um primeiro protótipo, desenvolvido no transcorrer deste trabalho, já conta com algumas funcionalidades que permiterm a sua operação e a obtenção de alguns resultados. Por ser incipiente, o sistema ainda carece de algumas melhorias na sua interface e na sua performance, mas principalmente é necessário realizar um conjunto significativo de testes a

fim de verificar a sua eficiência e eficácia. O sistema OntoSmart está acessível no endereço: **http://wrco.ccsa.ufpb.br:8080/ontosmart**.

Referências

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 2ª ed. Addison-Wesley, 2011.

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.

7
Considerações Finais

Diante da grande quantidade de informação disponível na Web, recursos de busca e recuperação de informação estão presentes em diversos *sites* para as mais variadas finalidades. Cotidianamente utilizamos tais recursos de forma natural e intuitiva para diversas tarefas rotineiras. Pesquisamos o menor preço de um determinado produto antes de adquiri-lo em uma loja virtual, que por sua vez possui um sistema que auxilia seus clientes na tarefa de encontrar o produto desejado. Nas livrarias *on-line* é possível encontrar obras do nosso autor favorito ou o *best-seller* do momento. Em *sites* corporativos é muito comum um campo busca com o qual podemos encontramos informações sobre um assunto de nosso interesse no contexto daquela empresa.

Essas ferramentas ou sistemas apresentam resultados relativamente satisfatórios, pois foram criados para atender a um domínio bastante restrito, no qual os itens de informação são conhecidos e as buscas podem ser facilmente previsíveis. Nesses "ambientes controlados" os problemas linguísticos são minimizados por permitirem a utilização de uma terminologia cujo campo semântico está restrito a uma determinada área, um assunto, ou mesmo a um ramo de atividade.

As ferramentas ou mecanismos de busca de propósito geral, tais como o Google, Yahoo!, AltaVista, Bing, têm pretensões universalistas de abarcar toda a informação livremente disponível na Web. A ausência de uma delimitação explícita do contexto semântico dos termos com os quais os documentos e as necessidades do usuário são representados afetam negativamente na precisão dos resultados de busca, que geralmente apresentam uma excessiva quantidade de documentos não relevantes (alta revocação).

Em um sistema de recuperação de informação existem dois elementos de natureza linguística: a *representação dos documentos* e a *representação da expressão de busca*. A eficiência do sistema é dependente da correta interpretação dos documentos e das necessidades de informação do usuário a fim de gerar suas respectivas representações. Além dos aspectos semânticos envolvidos nesse processo, tais representações devem estar formalmente estruturadas para que possam ser utilizadas por um sistema computacional.

Neste trabalho, os 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.

Como produto das pesquisas desenvolvidas no transcorrer deste trabalho, está em desenvolvimento o sistema denominado OntoSmart, cujo funcionamento foi detalhado no Capítulo 6. Embora ainda em processo de implementação, o sistema já apresenta alguns resultados que permitem tecer alguns comentários sobre suas vantagens e limitações.

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, 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 pelo usuário. 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.

O sistema OntoSmart está em desenvolvimento, não sendo possível ainda analizar de forma consistente o seu desempenho. Acredito, contudo, que ele venha a ser uma pequena semente, uma base para o desenvolvimento de futors trabalhos relacionados à recuperação de informação. Um primeiro tema que pode ser apresentado como uma possibilidade de pesquisa futura é a utilização de ontologias como interface de busca. Diversos editores e visualizadores de ontologias já contam com alguns recursos de aparesentação gráfica. Se adaptadas para o contexto da recuperação de informação, tais recursos poderiam auxiliar os usuários na elaboração de suas extratégias de buscas. A representação gráfica de ontologias poderia ser

utilizada também na apresentações dos documentos resultantes de uma busca, utilizando os conceitos como pontos centrais de agrupamentos de documentos (*clustering*).

Inicialmente, a definição dos *corpora* no sistema OntoSmart é realizado explicitamente por meio da definição dos documentos que os compõem e associando-lhes uma ontologia de domínio. Assim, um *corpus* é composto por um conjunto de documentos que versam sobre o domínio da ontologia a ele associado. Uma possiblidade para futuros desenvolvimentos está em criar agentes de software personalizáveis para coletar na Web documentos sobre um determinado assunto ou domínio específico e incluí-los no *corpus* de domínio correspondente.

Ao finalizar este trabalho, me surpreendo ao verificar o que foi realizado nesses exíguios 8 meses de estágio pós-doutoral. Foram quase uma centena de páginas de texto, resultado de uma pesquisa bibliográfica relativamente ampla, além do desenvolvimento de um sistema Web, o OntoSmart, um sistema ainda embrionário, mas que servirá de base para futuras pesquisas.

João Pessoa - PB
Praia de Cabo Branco