XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2013)
GT 8: Informação e Tecnologia

Comunicação Oral

UM MÉTODO DE EXPANSÃO AUTOMÁTICA DE CONSULTA BASEADA EM ONTOLOGIA

Edberto Ferneda – UNESP/MARÍLIA
Guilherme Ataíde Dias –UFPB

Resumo
A eficiência de um sistema de recuperação de informação depende da linguagem de representação dos itens de informação e das buscas dos usuários. O usuário procura traduzir a sua necessidade de informação em uma expressão de busca (consulta) a fim de recuperar documentos relevantes e úteis. A eficiência dessa tarefa é dependente do conhecimento da terminologia relacionada ao tema ou assunto de seu interesse. O processo de expansão visa melhorar a consulta inicialmente formulada pelo usuário agregando-lhe novos termos a fim de aumentar a precisão dos resultados obtidos. Este trabalho apresenta um método de expansão automática de consulta que utiliza ontologia como um vocabulário de domínio provedor de termos a serem utilizados na expansão da consulta inicialmente formulada pelo usuário. Considerando uma ontologia como um conjunto de conceitos interligados por relações semânticas, o método de expansão de consulta proposto neste trabalho considera que a distância entre dois conceitos de uma ontologia, dada pelo número de relações que os separa, reflete a proximidade semântica entre eles. Os termos de expansão, candidatos a integrar a consulta, são aqueles que mais se aproximam dos termos inicialmente utilizados na consulta do usuário. Os testes realizados atestam a efetividade do método, resultando em consultas mais específicas e, consequentemente, resultados mais precisos.
Palavras-chave: Expansão de Consulta. Ontologia. Recuperação de Informação.

A METHOD FOR ONTOLOGY-BASED AUTOMATIC QUERY EXPANSION

Abstract
The efficiency of an information retrieval system relies on the representation language of information items and the users' queries. The user tries to translate their information needs into a search expression (query) to retrieve relevant and useful documents. The efficiency of this task is dependent on the knowledge of the terminology related to the subject of interest. The expansion process aims to improve the query initially formulated by the user adding new terms to increase the results precision. This paper presents an automatic query expansion method using ontology as a domain vocabulary providing terms to be used in the query expansion previously formulated by the user. Considering an ontology as a set of concepts linked by semantic relations, the query expansion method proposed in this paper considers that the distance between two ontology concepts, given by the number of relations that separate them, reflects the semantic proximity between the concepts. The expression terms candidates to integrate the query, are those that most closely match the terms initially used in the user's query. The tests conducted evidenced the effectiveness of the method, resulting in more specific queries, and therefore more accurate results.
Keywords: Query Expansion. Ontology. Information Retrieval.

1 INTRODUÇÃO

Um sistema de recuperação de informação é um ambiente linguístico mediador da comunicação entre um estoque de informação e os seus requisitantes. Sua eficiência depende de um controle adequado da linguagem de representação dos itens de informação e das buscas dos usuários. Por meio de uma expressão de busca (consulta)[1] o usuário comunica a sua necessidade de informação e obtém como resultado um conjunto de documentos que possivelmente irão satisfazer tal necessidade. A eficiência em traduzir uma necessidade de informação em uma expressão de busca adequada dependente do conhecimento do usuário sobre a terminologia ligada ao tema ou assunto de seu interesse.

A importância e as dificuldades relativas à especificação de buscas fez surgir no interior da 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 aumentar a precisão dos 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.

Spink *et al* (2001) realizaram estudos envolvendo mais de um milhão de consultas realizadas na ferramentas de busca Excite[2] em um único dia: 16 de setembro de 1997. Observou-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. Constata-se, portanto, que as consultas inicialmente formuladas muitas vezes não resultam em um conjunto de documentos satisfatórios para as necessidades de informação dos usuários.

Este trabalho apresenta um método de expansão automática de consulta baseado em ontologia que pode ser aplicado em sistemas de recuperação de informação ou ferramentas de busca Web. O método aqui apresentado pode ser utilizado em sistemas booleanos, nos quais os termos de busca não possuem pesos, assim como em sistemas derivados do Modelo Vetorial, nos quais a cada termo de busca é associado um peso.

---

[1] Neste trabalho os termos "expressão de busca" e "consulta" serão utilizados como sinônimos. O primeiro é utilizado mais comumente na Ciência da Informação; o segundo é tradicionalmente utilizado na literatura da Ciência da Computação.

[2] www.excite.com

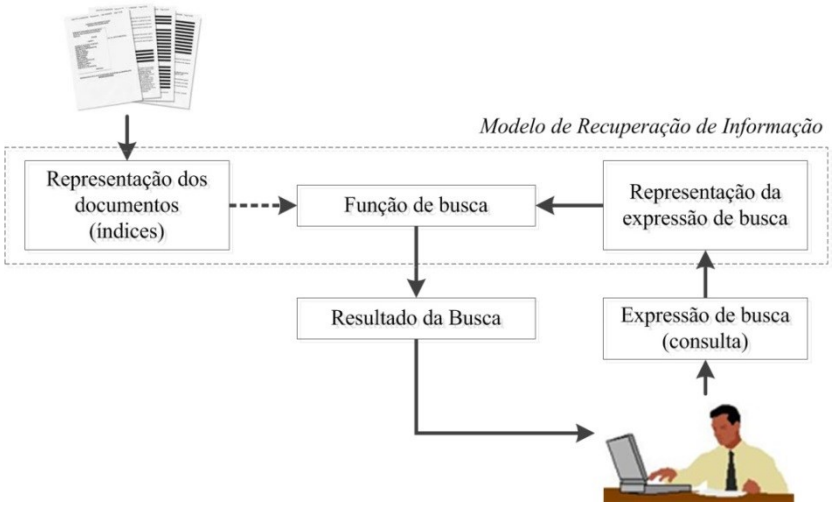
Considerando uma ontologia como um vocabulário controlado que contém a terminologia de uma determinada área do conhecimento (domínio), o método consiste em agregar termos derivados de uma ontologia à uma consulta inicialmente formulada pelo usuário com o objetivo de melhorar a eficiência do processo de recuperação de informação.

Na seção 2 será apresentada a área de Recuperação de Informação (*Information Retrieval*), o processo e os modelos clássicos de recuperação de informação. Na seção 3 define-se o conceito de ontologia e sua relação com a Ciência da Informação por meio da relação de semelhança com as linguagens documentárias. O conceito e os métodos de expansão de consulta são apresentados na seção 4. Na seção 5, após uma breve revisão da literatura sobre o tema, serão apresentados os conceitos básicos utilizados no método aqui proposto. Na seção 6 são apresentados alguns resultados experimentais e, por fim, na seção 7 serão expostas as conclusões e considerações finais deste trabalho.

2 RECUPERAÇÃO DE INFORMAÇÃO

O processo de recuperação de informação (Figura 1) consiste em identificar em um conjunto de documentos (*corpus*) de um sistema quais aqueles que atendem à necessidade de informação do usuário. Por um lado, para que um documento possa ser localizado e recuperado é preciso que ele tenha sido eficientemente representado por meio de um conjunto de termos de indexação. Por outro lado, a eficácia do processo de recuperação de informação está relacionada à representação linguística da necessidade de informação do usuário por meio da utilização de um conjunto de termos de busca.

Figura 1 - Representação do processo de recuperação de informação



Fonte: Adaptado de FERNEDA, 2012, p.14

No centro do processo de recuperação de informação está a *função de busca*, que compara as representações dos documentos com a representação da busca e recupera os

documentos que supostamente fornecerão ao usuário a informação que necessita. O resultado de uma busca é composto, geralmente, por uma lista de documentos em ordem decrescente de relevância, calculada pela função de busca.

A especificação formal da *representação dos documentos*, da *expressão de busca* e da *função de busca* compõe um *Modelo de Recuperação de Informação*. Os principais modelos de recuperação de informação foram criados entre as décadas de 1960 e 1980. Porém, as ideias e conceitos subjacentes a eles ainda estão presentes na maioria dos atuais sistemas de recuperação e nos mecanismos de busca da Web.

Os chamados "modelos clássicos" de recuperação de informação são propostas que serviram de base para o desenvolvimento de diversos outros modelos e algumas técnicas que até hoje são utilizadas. São eles: modelo booleano, modelo espaço vetorial e modelo probabilístico.

No *modelo booleano*, um documento é representado por um conjunto de palavras-chave ou termos de indexação. Uma busca é formulada por meio de uma expressão booleana composta por termos conectados através dos operadores lógicos AND, OR e NOT e apresenta como resultado os documentos cuja representação satisfaz às restrições lógicas da expressão de busca.

Uma das maiores limitações do modelo booleano está na impossibilidade de se ordenar (ranquear) os documentos resultantes de uma busca. Mas apesar de suas limitações, o modelo booleano está presente em quase todos os sistemas de recuperação de informação, seja como a principal maneira de formular as expressões de busca, seja como um recurso alternativo.

No *modelo vetorial* (SALTON; WONG; YANG, 1975) um documento é representado por um vetor onde cada elemento representa a relevância do respectivo termo na representação do seu conteúdo informacional. Uma expressão de busca é também representada por um vetor numérico onde cada elemento designa a importância do respectivo termo na representação da necessidade de informação do usuário (Figura 2).

Figura 2 – Exemplo de representação vetorial de um documento e uma expressão de busca

Fonte: Elaborada pelos autores

A utilização de uma mesma representação, tanto para os documentos como para as expressões de busca, permite calcular um valor numérico que representa o grau de

similaridade entre a expressão de busca e cada um dos documentos do *corpus*.Tais valores são utilizados no ordenamento dos documentos resultantes de uma busca.

O *modelo probabilístico* (ROBERTSON; JONES, 1976) trata 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 determinada expressão de busca. Assim, considera-se que 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 que representa a provável relevância do documento para a busca. Esse valor é utilizado para ordenar os documentos resultantes. Tendo um primeiro conjunto de documentos, o usuário pode marca alguns deles que considere verdadeiramente relevantes para a sua necessidade. O conjunto de documentos marcados pode então ser submetido ao sistema, fornecendo novos subsídios para um novo cálculo do grau de relevância de cada documento do *corpus*, o que permite fornecer resultados mais precisos. Esse processo interativo, 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.

Analisando-os em suas propostas originais, os três modelos clássicos consideram o processo de recuperação de informação como sistema fechado no qual o significado dos elementos lexicais é dado pelas suas inter-relações no interior de um *corpus* documental. Neste trabalho as ontologias são vistas como uma forma de vocabulário controlado cujos termos (conceitos) podem enriquecer as representações das buscas, proporcionando uma melhoria nos resultados de um sistema de recuperação de informação.

3 ONTOLOGIAS NA RECUPERAÇÃO DE INFORMAÇÃO

No seu papel de mediador de um processo comunicativo, é tarefa de um sistema de recuperação de informação 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 proporcionam uma convergência entre a linguagem do indexador e a linguagem do usuário de um sistema de informação. Segundo Tálamo, Lara e Kobashi (1992, p.197):

> 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. 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 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, 2009; JIMÉNEZ, 2004;). Dentre as semelhanças, pode-se destacar que: (1) ambos têm como objetivo representar e compartilhar os conceitos ou o vocabulário de um domínio a fim de possibilitar uma comunicação eficiente; (2) as suas estruturas básicas são hierárquicas, agrupando termos ou conceitos em categorias e subcategorias (classes e subclasses); (3) ambas podem ser utilizadas para catalogar ou organizar recursos informacionais. 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.

Pode-se definir uma ontologia como uma "especificação formal e 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 restrições e os axiomas devem estar formalmente definidos e passíveis de serem manipulados por computadores. Por *conceitualização* entende-se que tal especificação 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).

Para Jasper e Uschold (1999):

> 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.

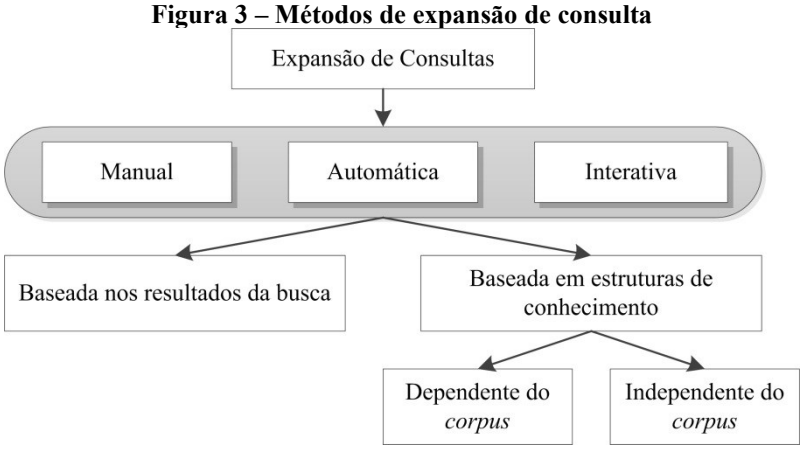
As ontologias se apresentam como um novo instrumento a ser incorporado ao arsenal teórico e prático da Ciência da Informação. A aprendizagem de novos conceitos e novos recursos oferecidos pelas ontologias é um desafio para os profissionais da informação, mas que pode ser facilmente enfrentado utilizando todo o conhecimento teórico e prático acumulado durante a história da Ciência da Informação.

4 EXPANSÃO DE CONSULTA

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 consequente aumento na revocação (*recall*). O operador AND restringe o resultado da consulta inicial, resultando em uma maior precisão. Nas abordagens baseadas no Modelo Vetorial, termos de expansão são adicionados à consulta original, juntamente com seus respectivos pesos (ROCCHIO, 1971).

Efthimiadis (1996) distingue três modos de expansão de consulta, como representado na Figura 3. Uma reformulação é considerada *manual* 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 adiciona à consulta original sem influencia ou mesmo sem a ciência do usuário. No modo *interativo* o usuário seleciona os termos de expansão a partir de um conjunto de termos apresentados pelo sistema.

Figura 3 – 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 3, estes termos podem ser originados dos *resultados de busca* ou de *estruturas de conhecimento*. Os métodos baseados nos resultados da busca selecionam os termos para expansão a partir dos documentos resultantes de uma consulta. Nesse caso, a eficácia da expansão da consulta depende fortemente da qualidade dessa consulta inicial. 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 poderão ser utilizados na expansão da consulta. Mecanismos independentes do *corpus* contam com estruturas de conhecimento que não apresentam relação com os documentos. São exemplos dessas estruturas: léxicos, glossários, dicionários, tesauros, ontologias. 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, caso em que os mecanismos dependentes do *corpus* não serão muito eficazes.A aplicabilidade dos métodos de expansão de consulta baseados em estruturas independentes do *corpus*, por outro lado, independem da quantidade de documentos ou da extensão destes.

Outra vantagem do uso de estruturas de conhecimento na expansão da consulta é sua disponibilidade a qualquer momento no processo de busca. A formulação da primeira consulta do usuário já pode se beneficiar deste tipo de expansão, pois os termos não são derivados de resultados da uma busca. 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 em um impulso significativo a esse tipo expansão de consultas.

5 UM MÉTODO DE EXPANSÃO DE CONSULTAS BASEADA EM ONTOLOGIA

Será descrito nesta seção alguns conceitos básicos do método de expansão de consulta proposto e a sua forma de utilização. Conforme a classificação de Efthimiadis (1996), trata-se de um método de expansão automática baseada em estrutura de conhecimento (ontologia) independente do *corpus*.

Dey *et al*. (2005) implementaram um mecanismo de expansão de consulta no qual a determinação das condições de expansão é calculada a partir da distância semântica entre os termos da consulta e os conceitos de duas ontologias: uma ontologia sobre vinhos e outra sobre plantas. Como resultado de suas experiências utilizando o Google, os autores relatam um aumento na precisão das consultas que foram expandidas com o uso dos termos das ontologias.

Sack (2005) demonstrou que 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 bibliográficos e uma ontologia do domínio de um tipo de problemas computacionais denominados “NP-completos”. Essa ontologia foi utilizada na expansão de consultas 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 do domínio de interesse.

O sistema FOQuE (YAGUINUMA; BIAJIZ; SANTOS, 2007), utiliza ontologias difusas para recuperar resultados aproximados e semanticamente relevantes, de acordo com parâmetros de expansão definidos pelo usuário. As respostas adicionais são classificadas conforme o tipo de expansão realizada e a relevância para a consulta, permitindo, assim, que o usuário identifique quais os resultados aproximados mais apropriados para seus requisitos.

Elias (2010) apresenta um estudo sobre a utilização de ontologias de domínio e outras artefatos de controle terminológico para melhorar a eficiência na recuperação de informação. Utiliza o conhecimento de domínio presente nestes artefatos como fonte de termos relacionados para complementar a consulta inserida pelo usuário com o objetivo de desenvolver uma técnica de

expansão de consultas que melhore tanto a precisão quanto a cobertura de um sistema de recuperação de informação.

O método de expansão de consulta aqui proposto considera que a distância entre dois conceitos de uma ontologia reflete a proximidade semântica entre eles. Os termos de expansão, candidatos a integrar a consulta, são aqueles que mais se aproximam dos termos inicialmente utilizados na consulta do usuário.

5.1 DISTÂNCIA SEMÂNTICA (*DS*)

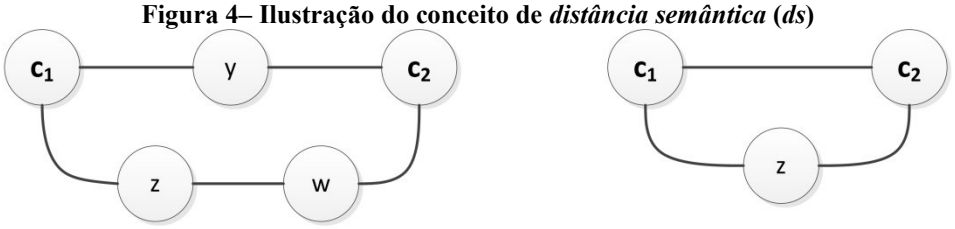
Uma ontologia O = (C, R) é composta por um conjunto de conceitos C={c_1, c_2,.., c_n} interconectados por um conjunto de relacionamentos em R={r_1, r_2, ..., r_n}. Define-se inicialmente a *distância semântica* entre dois conceitos:

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**.*

No exemplo da Figura 4(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 4(b) *ds*(c_1, c_2)=1, pois os conceitos **c_1** e **c_2** são adjacentes, separados por um único relacionamento.

Figura 4– Ilustração do conceito de *distância semântica* (*ds*)



(a) **(b)**

Fonte: elaborada pelos autores

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

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

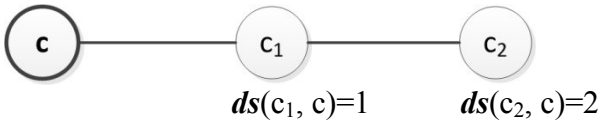
Definição 2

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

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

onde **p** *é um parâmetro numérico, entre 0 e 1, que define a diferença dos valores de* **vs** *a cada distância* **ds**.

Na figura abaixo temos um conceito central **c** relacionado aos conceitos $\mathbf{c_1}$ e $\mathbf{c_2}$. A distância semântica de $\mathbf{c_1}$ em relação a **c** é igual a 1. A distância semântica de $\mathbf{c_2}$ em relação a **c** é igual a 2.



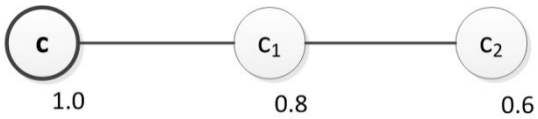
Considerando o parâmetro $\boldsymbol{p}$ = 0.2, pode-se calcular o valor semântico (***vs***) dos conceitos $\mathbf{c_1}$ e $\mathbf{c_2}$ em relação à **c** como:

$$\boldsymbol{vs}(c_1, c) = 1 - [\boldsymbol{ds}(c_1, c) \times 0.2] = \mathbf{0.8}$$

$$\boldsymbol{vs}(c_2, c) = 1 - [\boldsymbol{ds}(c_2, c) \times 0.2] = \mathbf{0.6}$$

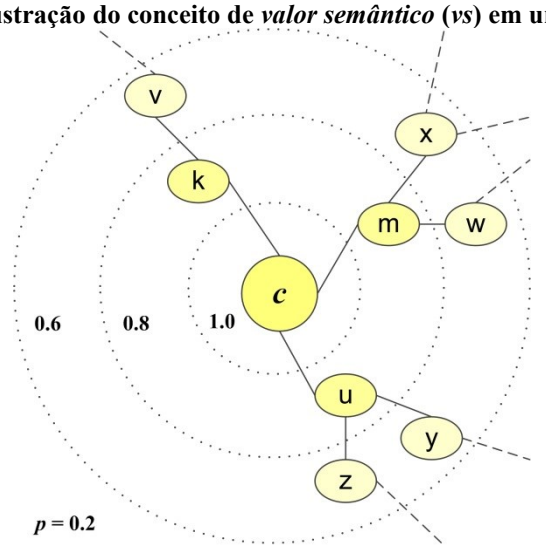
Como apresentado anteriormente, a distância semântica (***ds***) de um conceito em relação a ele próprio é igual a zero. Assim, o valor semântico (***vs***) de um conceito em relação a ele mesmo é igual a 1.

Utilizando o conceito **c** como referência (conceito central) e $\boldsymbol{p}$= 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 5 apresenta uma ilustração mais ampla da aplicação dos conceitos de distância semântica (***ds***) e valor semântico (***vs***), considerando o parâmetro ***p***=0.2.

Figura 5– Ilustração do conceito de *valor semântico* (*vs*) em uma ontologia



Fonte: elaborada pelos autores

A definição de um conceito central em uma ontologia faz surgir diversos níveis ou "camadas" concêntricas, onde cada camada é definida pela distância semântica (***ds***) em relação a um 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 conforme o valor do parâmetro ***p***. Considerando **c** o conceito central e ***p***=0.2, os conceitos da Figura 5 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 5 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 novo parâmetro ***k*** que limite o número de camadas a serem consideradas no cálculo de ***vs***.

Os valores dos parâmetros ***p*** e ***k*** são interdependentes. Não faz sentido, por exemplo, ***p***=0.2 e ***k***=8, pois isso acarretaria valores negativos de ***vs***. Portanto, o valor máximo que o parâmetro ***p*** pode assumir ($p_{\max}$) é igual 1/***k***. No exemplo da Figura 5 o valor de ***k*** é igual a três (***k***=3). Portanto, o valor máximo que do parâmetro ***p*** pode assumir é 0.33 (p_{max}=0.33).

Como visto, tendo como referência um conceito central, é possível definir "camadas" concêntricas entre este e os demais conceitos de uma ontologia utilizando-se a medida ***ds***. A partir do conceito central, cujo ***vs*** é igual a 1, é possível ainda atribuir aos demais conceitos valores decrescentes de ***vs*** à medida se afastem do conceito central. A diferença de ***vs*** a cada

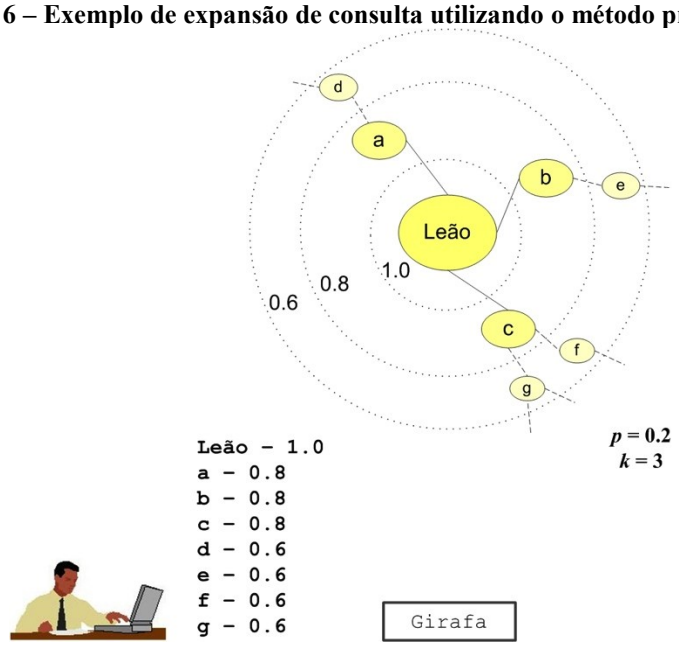
distância ***ds*** é dada pelo parâmetro ***p*** e a abrangência (número de camadas a ser considerado) é determinado pelo parâmetro ***k***.

5.3 O PROCESSO DE EXPANSÃO DE CONSULTA

Antes da execução qualquer consulta, o usuário deve selecionar a ontologia do domínio ao qual se refere a sua necessidade de informação e determinar valores para os parâmetros ***k*** e ***p***. Cada termo da consulta do usuário será utilizado como conceito central da ontologia. Assim, 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 utilizados na consulta e que se encontram diretamente representados na ontologia.

A Figura 6 ilustra 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.

Figura 6 – Exemplo de expansão de consulta utilizando o método proposto



Fonte: elaborada pelos autores

Tomando-se "Leão" como o conceito central da ontologia e considerando os parâmetros ***p*** = 0.2 e ***k*** = 3, derivam-se os termos **a**, **b, c** com o peso igual a 0.8 e **d, e, f**, **g** com peso 0.6. Em sistema baseados no modelo vetorial esses termos e seus respectivos pesos podem ser utilizados para formar o vetor de busca. Em sistemas que não atribuem pesos os termos da consulta, tal como no modelo booleano, esses valores podem ser ignorados ou utilizados para compor a expressão booleana da consulta.

O termo “Girafa” foi descartado por não estar representado por um conceito da ontologia. Termos que foram encontrados na ontologia serão armazenados em um tipo de repositório, formando um conjunto de potenciais conceitos, que podem vir a fazer parte da ontologia.

Uma consulta é formulada e submetida após a escolha da ontologia relacionada ao tema ou assunto de interesse do usuário. Assim, se um termo que **não** está presente em uma determinada ontologia for repetidamente utilizado nas consultas, pode-se supor que talvez ele deva ou deveria fazer parte dessa ontologia. Esse processo de povoamento de ontologias permite refletir a flexibilidade e dinamicidade inerente às linguagens humanas.

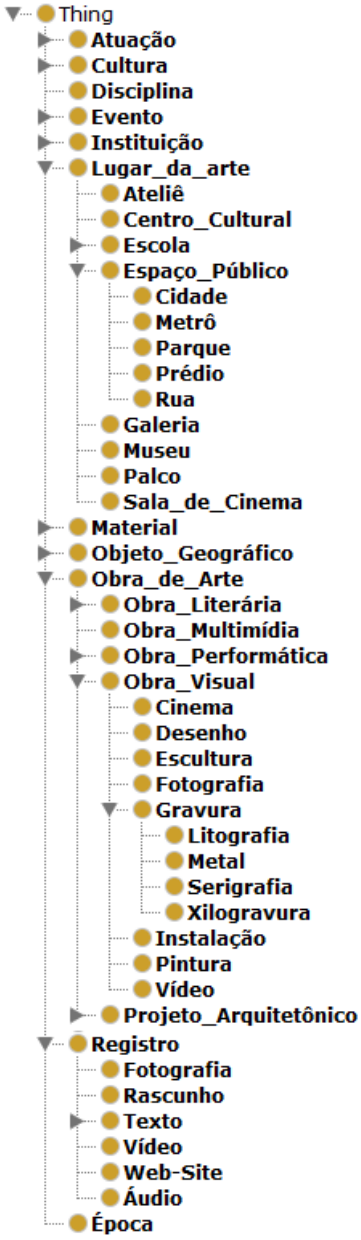
6 RESULTADOS EXPERIMENTAIS

O método de expansão de consulta aqui descrito foi implementado em um sistema de recuperação de informação denominado OntoSmart. Esse sistema foi desenvolvido pelos autores deste trabalho entre meados de 2012 e início de 2013 com o propósito de servir como base para pesquisas na área de Recuperação de Informação.

Para a realização dos testes, foi utilizada uma ontologia no domínio de Arte Contemporânea[3] disponível na Web. A Figura 7 mostra parte da hierarquia de classes dessa ontologia, apresentada no software Protégé.

[3] Disponível em http://www.ime.usp.br/~rmcobe/onair/files/arte-contemp.owl. Acessado em 06/08/2013

Figura 7 – Ontologia sobre Arte Contemporânea



Fonte: http://www.ime.usp.br/~rmcobe/onair/files/arte-contemp.owl
apresentada no Protégé 4.**3**

Uma limitação atual do sistema utilizado para os testes (OntoSmart) é sua restrição a um único termo na consulta inicial. Os testes foram conduzidos de forma direcionada, com a utilização de termos relacionados a conceitos sabidamente existentes na ontologia a fim de analisar a consulta resultante. Os resultados de algumas consultas são apresentados na Tabela 1.

Tabela 1 – Resultado dos testes realizados

Consulta inicial	Parâmetros		Consulta expandida
	p	***k***	
Crônica	0.1	2	Crônica – 1.0; Obra Literária – 0.9
Crônica	0.2	3	Crônica – 1.0; Obra Literária – 0.8; Texto – 0.6
Crônica	0.3	4	Crônica – 1.0; Obra Literária – 0.7; Texto – 0.4; Registro – 0.1
Xilogravura	0.2	2	Xilogravura – 1.0; Gravura – 0.8
Xilogravura	0.2	4	Xilogravura – 1.0; Gravura – 0.8; Obra Visual – 0.6; Obra de Arte – 0.4
Fotografia	0.2	3	Fotografia – 1.0; Obra Visual – 0.6; Registro – 0.6
Fotografia	0.4	2	Fotografia – 1.0; Obra Visual – 0.8; Registro – 0.8; Obra de Arte – 0.6
Metrô	0.2	3	Metrô – 1.0; Espaço Público – 0.8; Lugar da Arte – 0.6

Fonte: Elaborada pelos autores

Pode-se observar que o parâmetro ***k*** possui relação com a especificidade ou generalidade da consulta. Quanto maior o seu valor maior também será o número de termos da consulta. Em um sistema booleano, considerando que os termos serão ligados pelo operador AND, quanto maior o valor de ***k*** maior será a especificidade da consulta e, teoricamente, maior será a precisão do processo de recuperação.

O parâmetro ***p*** reflete a importância do termo na representação da necessidade de informação do usuário. O termo utilizado na consulta inicial, se existente na ontologia, possui peso unitário. Os termos adjacentes a ele terão pesos menores, variando conforme a distância (número de relações) que os separa daquele. O parâmetro ***p*** é notadamente útil em sistemas de recuperação baseados no modelo vetorial.

Com mencionado, antes de realizar sua consulta o usuário deve definir (selecionar) a ontologia referente ao assunto de interesse. Dessa forma restringe-se o campo semântico dos termos da consulta, permitindo que o sistema expanda a consulta de forma eficiente. Como apresentado na Tabela 1, a consulta que utiliza o termo "Crônica" foi expandida por meio da agregação do termo "Obra Literária", o que é condizente com o contexto ao qual se refere (Arte Contemporânea). À consulta inicial com o termo "Xilogravura" foi incluída o termo "Gravura", com ***k*** = 2. Com ***k*** = 4 foram acrescidos os termos "Obra Visual" e "Obra de Arte", o que contextualiza e especifica melhor a necessidade de informação do usuário. Da mesma forma, em um contexto de arte, o termo "Metrô" se refere a um "Espaço Público" que pode servir como um "Lugar da Arte".

7 CONCLUSÕES

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 dos usuários 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 (conceitos) são considerados termos de um vocabulário de domínio, utilizado como ferramenta para expansão das consultas em um sistema de recuperação de informação. A explicitação *a priori* do contexto da busca por meio da seleção da ontologia permite fornecer um conjunto de termos relevantes e qualificados, proporcionando eficácia na expansão da consulta inicialmente formulada.

Embora os testes realizados sejam ainda incipientes, eles já possibilitam comprovar a eficiência do método proposto. Os testes mostraram que a agregação de termos derivados de uma ontologia de domínio possibilita expandir a consulta inicial de forma eficiente, gerando expressões mais específicas, que resultariam em um aumento na precisão dos resultados da recuperação. Novos testes poderão ser realizados a partir do amadurecimento do sistema OntoSmart e da formação de um *corpus* documental significativo. Isso permitiria avaliar a eficiência desse método de expansão de consultas a partir da análise dos documentos resultantes da recuperação utilizando as medidas de precisão e revocação.

AGRADECIMENTOS

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

REFERÊNCIAS

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

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

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

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.

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

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

ELIAS, A.B. **Expansão semântica de consultas baseada em esquemas terminológicos**: uma experimentação no domínio biomédico. 2010. 131 f. Tese (Mestrado em Informática) – PPGI, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, 2010.

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

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

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.

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

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

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

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

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.

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

ROBERTSON, S.E.; JONES, K.S. Relevance weighting of search terms. **Journal of the Americam Society for Information Science**, v. 27, n. 3, 1976.

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.

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

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

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

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.

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.

SALTON, G. Experiments in Automatic Thesaurus Construction for Information Retrieval. In: FREIMAN, C. V.; GRIFFITH, J.E.; ROSENFELD, J.L. (eds.) **Information Processing 71**: Proceedings of IFIP Congress 71, v.1. North-Holland, 1972.

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

YAGUINUMA, C. A.; BIAJIZ, M.; SANTOS, M. T. P. Sistema FOQuE para Expansão Semântica de Consultas Baseada em Ontologias Difusas. In: XXII Simpósio Brasileiro de Banco de Dados, 2007, João Pessoa (PB). XXII Simpósio Brasileiro de Banco de Dados. Porto Alegre: SBC, 2007. v. 1. p. 208-222.