Expansão de consulta semântica aplicadas a Sistemas de Recuperação de Informação de contexto Geográfico

Leila Weitzel (Organizador)[1], José Palazzo Moreira de Oliveira (Orientador)[1], Joel Luis Carbonera[1], Paulo André Torres[1]

Resumo: O objetivo desta pesquisa é desenvolver uma ontologia que pudesse promover melhorias no desempenho de sistemas de recuperação de informação de contexto geográfico. A metodologia de desenvolvimento seguiu os parâmetros do Método 101. Para validação da ontologia propõe-se aplicar a técnica de expansão semântica (manual) das consultas e submetê-las ao sistema Lemur para verificação.

Abstract: The main goal of this research is to develop an ontology which could promote improvements in the performance of information retrieval systems under geographic context. The Ontology Engineering Methodology followed the Ontology Development 101. To validate the ontology it is considered to apply (manual) semantic query expansion and submit them to the Lemur system.

1 Introdução

Este relatório está no âmbito da disciplina CMP234 - Modelagem Conceitual e Ontologia do PPGC- UFRGS, Programa de Pós-Graduação em Computação ocorrida no segundo semestre de 2009. Durante esse período o aspecto que norteou a pesquisa foi o processo de construção de uma Ontologia que pudesse promover o desempenho da Recuperação de Informação de conteúdo Geográfico na Web. O desafio GeoCLEF 2008 - Cross-language Geographic Information Retrieval - foi o grande motivador desta pesquisa, pois a versão

1 Instituto de Informática, UFRGS, Caixa Postal 9999
{lwcsilva,palazzo,joel.carbonera, paulo.torres@inf.ufrgs.br}

CLEF de 2008 tem justamente como propósito avaliar sistemas de recuperação de informação de contexto geográfico na Wikipédia.

2 Ambientação da pesquisa

O volume de informação disponível na Web tem crescido nos últimos tempos em uma velocidade maior que a capacidade de processamento desta informação. O crescimento acelerado da quantidade de informação reflete: a rapidez do desenvolvimento de novas tecnologias de informação; o barateamento de tecnologias/dispositivos de aquisição de informações e a proliferação de portais sociais e de informação tais como youtube, facebook, twitter, wikipedia, flickr, blogs, etc. Este problema é constantemente referenciado na literatura como o problema da "sobrecarga de informações" ou "information overload", [1].

As primeiras soluções para o problema de gerenciamento do excesso de informação na Web foram dadas por ferramentas de busca baseadas em filtragem da informação através de meta-informações, com a utilização de palavras-chave além de heurísticas para categorização. Desde então, os princípios da Web Semântica têm sido pesquisados, aplicados, revistos e aprimorados por diversos grupos de trabalho [2], [3]. A Web tem demando sistemas capazes de reconhecer e tratar dados, possibilitando o compartilhamento em relação a significado de conceitos. Neste sentido, a web semântica representa um empreendimento que busca tratar as informações da Web como uma rede de conceitos em contraposição a uma rede de documentos, de modo que seja possível associar conhecimento do significado aos recursos da Web, tipicamente através da utilização de (meta) dados processáveis por máquinas [5].

Neste contexto, este trabalho aborda a questão da usabilidade de ontologias para promover o desempenho da Recuperação de Informação geograficamente contextualizadas na Web sob uma perspectiva multilíngüe. Assim, o foco do trabalho é estabelecer uma ponte semântica entre os documentos a serem recuperados e a ferramenta de busca, de modo que seja possível obter resultados da pesquisa que possam refletir de forma mais adequada a intenção e as conceitualizações do usuário.

Na ultima década a Ontologia vem compondo a espinha dorsal da Web Semântica [4], possibilitando o preenchimento do "vazio" semântico entre a representação sintática da informação e sua conceitualização. Uma ontologia é uma teoria que especifica um vocabulário relativo a certo domínio, definindo entidades, classes, propriedades, predicados e funções e as relações entre estes componentes; e tem como objetivo geral explicitar e compartilhar conhecimento comum sobre a estrutura da informação entre pessoas, sistemas e agentes de software [6]. Para o compartilhamento e reuso deste conhecimento faz-se necessária a elaboração de restrições entre os conceitos, de modo que o conhecimento armazenado se torne coerente e transparente. A ontologia para Web Semântica é um vocabulário que define conceitos e metadados, que são usados principalmente para compreender interoperabilidade semântica. Em [7], [8], [9], [10], [11], [12], [13], [14] encontram-se trabalhos precursores e do "estado da arte" em ontologia.

Na web, o volume, os diferentes idiomas e dialetos em que as páginas são escritas e as especificidades culturais dificultam a recuperação de informação, fazendo com que a recuperação de informações tradicionalmente realizada retorne uma grande quantidade de ruído. Desta forma, a eficiência de um sistema de Recuperação de Informação é avaliada pela sua capacidade em apresentar informações que atendam às necessidades dos usuários. As medidas de desempenho (precisão e revocação) são baseadas na noção de documentos relevantes, de acordo com uma determinada necessidade de informação. Somados a isso, tem-se que o contexto geográfico é transversal a inúmeros domínios do conhecimento (turismo, jurídico, meio ambiente, entre outros). Apresenta-se a abaixo um texto retirado do sítio do MMA – Ministério do Meio Ambiente notícia datada do dia: 30/11/2009:

"Um dos parques mais antigos do Brasil, o Parque Nacional da Serra dos Órgãos teve as comemorações de 70 anos de sua fundação nesta segunda-feira (30/11), em Teresópolis. Inaugurado em 1939, o local foi criado para proteger a paisagem e a biodiversidade do trecho da Serra do Mar na região serrana do Rio de Janeiro - consideradas por especialistas como excepcionais..."

Apenas neste trecho do texto é possível notar várias informações que remetem à noções geográficas:

(i) classe, objetos e instâncias geográficas,
(ii) contexto geográfico (região serrana)
(iii) informações espaço-temporais (ano da fundação do parque).

A semântica subjacente aos conceitos poderia ser: Teresópolis (Cidade) na Região Serrana (região de serra) que é um trecho da Serra do Mar (cadeia de montanhas que acompanha cerca de 1.000 km do litoral sul e sudeste do Brasil entre Rio de Janeiro e o norte de Santa Catarina), que pertence ao Estado do Rio de Janeiro (Um Estado do Brasil na Região Sudeste), possui um Parque Nacional da Serra dos Órgãos (Reserva Florestal), e todas as outras descrições do contexto geográfico presente no texto.

3 Processo de construção da ontologia

3.3 Considerações da pesquisa

De acordo com Fernandez-López & Gómez-Peréz [15] não existe uma metodologia completamente madura para o propósito de construção de ontologias. Segundo os autores, uma combinação de metodologias se torna interessante em todo o processo. Sendo assim o processo de construção da Onto-GeoCLEF foi baseado na Metodologia 101 [14] e no On-toKnowledge [4]. A Metodologia 101 propõe que o processo de construção deva ser iterativo e com sete passos: determinar o escopo da ontologia, considerar o reuso, listar termos, definir classes, definir propriedades, definir restrições e criar instâncias, abriga as questões de competência permitindo um modo simples e direto para se determinar o escopo da ontologia [14]. Já a On-to-Knowledge [4] é uma metodologia de desenvolvimento de ontologias fruto da cooperação de várias entidades européias e que tem como propósito o desenvolvimento de

ontologias para serem empregadas em Sistemas de Gestão do Conhecimento. Esta metodologia é baseada em cinco fases:

(i) Estudo da viabilidade;

(ii) Inicio da ontologia (definição do domínio, objetivos, etc.);

(iii) Refinamento (técnicas de elicitação de conhecimento);

(iv) Avaliação e

(v) Manutenção e evolução.

De cada uma das metodologias empregadas no trabalho utilizou-se diferentes fases de cada uma delas.

A primeira etapa na construção da Onto-GeoCLEF foi baseada no primeiro passo da Metodologia 101 [14] e na Segunda fase da On-to-Knowledge [4] Nesta etapa foram especificados: os requisitos, definindo o domínio e objetivos da ontologia, enumerando questões de competência, definindo o ambiente de desenvolvimento da ontologia, entre outros.

O Geo-CLEF-2009 tem como propósito avaliar sistemas de recuperação de informação multilíngüe de contexto geográfico. Desta forma, baseado no empirismo, levantou-se o conjunto de conceitos que pudessem expressar a semântica dos Termos para que se obtivesse a mais correta interpretação das consultas feitas por um utilizador aos sistemas buscadores.

De acordo com [14], as questões de competências são relevantes no processo de construção de ontologias. Essas questões devem refletir de forma mais adequada a intenção e as conceitualizações requeridas pelo sistema de busca. As principais questões que tiveram lugar foram:

a) Documentos com conteúdo geográfico que apresentem relações de vizinhança, incidência e sobreposição (perto de, dentro de, etc). O utilizador pode estar querendo obter informações de determinado local em relação a posição de outro. Ou ainda, obter informação de locais em relação à sua posição cardinal (ao norte, ao sul, etc).

Consulta: Quais são os restaurantes japoneses ao norte da Ilha de Manhattan?

b) Documentos que apresentem relações entre conceitos geográficos e eventos (ou ocorrências).

Consulta: Onde e quando foi encontrado "Lucy"? Onde ocorreu o mais antigo e devastador Tsunami?

c) Documentos que apresentem relações entre dois ou mais eventos.

Consulta: Existe outro achado fóssil mais ao sul de onde foi encontrado "Lucy"?

d) Documentos que contenham adjetivos pátrios ou gentílicos. Os gentílicos exprimem a procedência ou a naturalidade. É a forma de designar alguém ou algum objeto em função do país, da região, da província, da localidade em que nasceu ou de onde alguém ou alguma coisa procede.

Consulta: Existe algum bar típico Carioca em São Paulo? Quais são os mais antigos times de futebol Paulista?

e) Documentos que se referem a locais que possuem coordenadas geográficas, mas que historicamente não são mais mapeados. Por exemplo, A cidade de São Sebastião do Rio de Janeiro era a capital do Brasil colonial e hoje é capital do estado do Rio de Janeiro. Outro exemplo, Alemanha oriental e ocidental unificadas forma o atual País Alemanha.

f) Documentos que podem se referir tanto a uma pessoa natural quanto a um conceito geográfico. Ou seja, alguns Termos podem designar objetos diferentes, ou Termos diferentes podendo designar o mesmo objeto.

Consulta: Quais são as façanhas de Bento Gonçalves? Ou Quais os vinhos de Bento Gonçalves? Na primeira, o usuário está buscando informações sobre a participação do Republicano na Revolução Farroupilha, onde Bento Gonçalves é instância do conceito Personagens Históricos e na segunda uma instância de Cidade que é um objeto geográfico.

g) Documentos que façam referência a Pessoa Natural ou papel que ela desempenha em determinado intervalo de tempo, ou lugar.

Consulta: Quais são presidentes que tiveram barba na historia mundial? Neste caso são pessoas naturais que estão desempenhando papel de Presidente, ou Primeiro Ministro, Estudante, Músico, Cantor, etc.

h) Documentos que contenham referência a uma pessoa ou a um objeto pela sua antonomásia.

Consulta: Qual a data de nascimento do Rei do Rock-and-Roll? (Elvis Presley). Qual o nome e ano da primeira novela da Namoradinha do Brasil? (Regina Duarte). Ano de nascimento do Rei do Futebol? (Edson Arantes do Nascimento). Qual a local de origem do Patrono do Exercito Brasileiro (Luís Alves de Lima e Silva)

i) Documentos que se refiram a objetos, lugares ou pessoas pelo seu apelido ou seu título.

Consulta: Qual o ano de morte de Tiradentes? Qual a cidade e o ano de nascimento do Lula? Qual o nome da filha da Xuxa? (Maria das Graças Meneghel) A Cidade natal do Xuxa? Neste caso é o nadador Fernando de Queiróz Scherer.

3.4 Descrição semântica dos termos

A questão semântica no âmbito da Ciência da Informação tem se apoiado nas teorias relacionadas à representação de sistemas de conceito, tais como: Teoria da Classificação Facetada de Ranganathan [17] e a Teoria de Conceito de Dahlberg [16].

A Teoria de Conceito de Dahlberg [16] foi desenvolvida no campo da elaboração de Tesauros. Para o autor um objeto individual representa um CONCEITO. O CONCEITO é uma Unidade de Conhecimento, de tal forma que para cada CONCEITO, um existe somente um conjunto particular de enunciado. Implica que para cada CONCEITO tem-se um referente (seja este um conjunto de objetos, um único objeto, uma atividade, um fato, um tópico, etc.), sobre o qual afirmações verificáveis podem ser feitas. Essas afirmações podem ser sumarizadas e/ou sintetizadas por um TERMO que, então, representará um CONCEITO. O REFERENTE é aquilo que se pretende conceituar, as CARACTERÍSTICAS são as soma dos enunciados verdadeiros sobre este Referente. Desta forma, os termos que foram determinados no primeiro passo da Metodologia 101 serão descritos utilizando-se dos preceitos de Dahlberg [16].

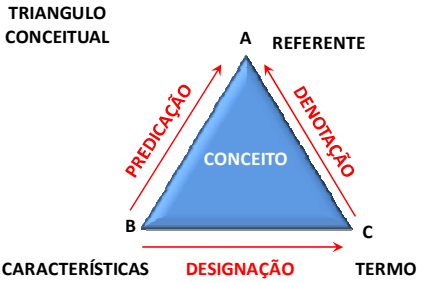

Figura 1. Triangulo Conceitual de Dahlberg (1978)

A = Item de Referência (IR)
B = Predicações Verdadeiras (PV) sobre IR
C = Síntese das PV sobre IR, por meio de um Termo/nome.

Nesta pesquisa também se utilizou a abordagem sobre os relacionamentos semânticos que seriam apropriados às ontologias, pois acredita-se que estes relacionamento poderiam melhorar os processos de recuperação da informação de contexto geográfico a Web.

As Relações Semânticas Monolíngües que foram utilizadas são as de: HIERARQUIA, INCLUSÃO, EQUIVALÊNCIA E OPOSIÇÃO. Classificandas em três de relações binárias: a relação classe-classe (a relação "é um", is-a), a relação instância-classe ("é instância de", instance of) e a relação instância-instância (por exemplo, a relação "é parte de", part-of "é equivalente a" equivalent-to).

As Relações Semânticas Multilíngües propostas foram: sinônimos-totais que relaciona conceitos de diferentes línguas que podem ser livremente substituídos (City=Cidade); A relação equivalente-hiponímia e equivalente-hiperonímia, que relaciona um conceito mais específico em uma língua a um conceito mais geral em outra língua, e vive-versa, por exemplo, "neve" em português do Brasil, Neige em Frances ou Snow em Inglês.

Com base nesse primeiro estágio de estudo do vocabulário do domínio e suas relações, passamos a nos envolver com a descrição explícita do conhecimento semântico, ou seja, na descrição dos termos levantados.

Termo 1: GEOMETRY
Definição: Adotou-se a definição do termo usado em Sistemas de Informações Geograficas: termo geralmente usado para descrever a maneira como um objeto do mundo real "Entidade" é representado geometricamente em um banco de dados. Geometria é usada para a representação da componente espacial de uma feição geográfica, pode ser do tipo ponto, polígono ou linha.
Características:
Data Properties: NomeGeo: range (polígono, linha e ponto)
Termo 2: SPECIALNAMES

Definição: é um conceito relacionado a estratégias retóricas ou a figures de linguagem, isto é, utiliza-se parte do nome de um Agent ou Place indicando seu vínculo familiar; sobrenome, nome de família ou designação informal para identificar uma pessoa, objeto ou lugar; alcunha, apodo.

Características:

Data Properties: NameSpecial (string)

Subclass

1. **Nicknames**: é um nome de **agent** ou **place** pelo qual estes podem ser conhecidos informalmente. Exemplo: Tiradentes

Antonomasia: uma figura de linguagem caracterizada pela substituição por um nome de uma expressão que lembre uma qualidade, característica ou fato que de alguma forma identifique-o. Exemplo Rei do Futebol, Namoradinha do Brasil

Termo 3: GENTILIC

Definição: Descreve um P**lace**. Exemplo: carioca, gaucho, fluminense.

Características:

Data Properties: GentilicName (string)

Object Properties: HasIdiom

Termo 4: PLACE

Definição: Define uma certa localização.

Características:

Object Properties: HasSpecialNames, HasGeometry, HasName, HasPopulation, HasTime, HasIdiom, HasGentilic, HasPart (atomic part, proper part: Boundary, Temporary-proper-part), IsParticipantOf (temporal participant, constant participant), HasTopological (disjoint, meet, overlap, contains, equal, cover-by, inside, covers), HasMetric (Metrics: distance and others), HasCardinal (north, south, Etc...)

Subclass:

a) **Absolute Place**: qualquer região no espaço dimensional que é utilizado para localizar uma "Entidade". É um lugar em um senso genérico que possui uma coordenada geográfica. Por exemplo, um país, estado, logradouro, cidade etc.

b) **Relative Place**: qualquer região espaço-temporal. Por exemplo, a Cidade São Sebastião d Rio de Janeiro é a cidade do Rio de Janeiro nos tempos atuais. Alemanha Oriental e Alemanha Ocidental é hoje em dia a Alemanha, assim Alemanha é um AbsolutePlace e Alemanha Oriental e Ocidental é seu RelativePlace.

Subclass:

i. **Historical Relative Place**: Alemanha Ocidental

ii. **Region Relative Place:** Terceiro Mundo, Região dos Bálcãs, etc.

Termo 6: AGENT

Definição: É um objeto "Entidade" física ou abstrata, indivíduos ou grupo de indivíduos.
Características:

SubClass

1. **Legal person**: é uma pessoa legal perante as leis.
2. **Person**: é uma pessoa natural no senso comum.

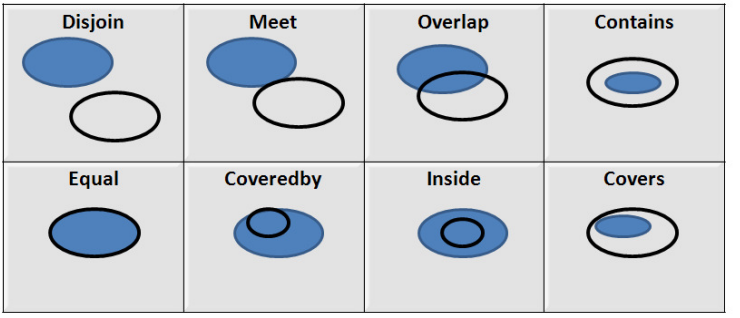

Figura 2. Relações topológicas no espaço $\mathbb{R}^2$ (Egenhofer & Franzosa, 1991)

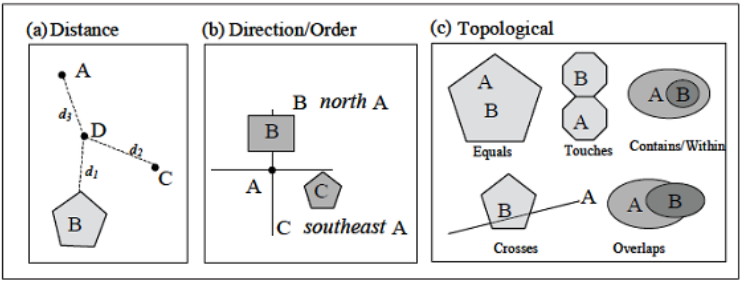

Figura 3. Relações espaciais

Termo 7: ROLE
Definição: Papel é desempenhado por um **Agent** em um contexto social, exemplo advogado, Músico, etc.

Termo 8: Artwork:
Definição: manifestações artísticas em geral
Subclass:

- Sculpture
- Music

- Dance
- Movie
- Literature
- Theater
- Paint
- Etc...

Termo 9: Occurrence
Definição: Ocorrências ou eventos compreendem um fenômeno um acontecimento que pode ser natural ou artificial.
Características:
Subclass:
i. **Natural events**:
Subclass
a) **Natural Disasters;** enchentes, terremotos, etc.
b) **Artificial disasters**: guerras, etc.
ii. **Human Activities**: encontros, concertos de rock, descobertas, olimpíadas, jogos de inverno etc.

Termo 10: Name
Definição: designa e identifica **Agent** e **Place**.
Características:
Object Properties: HasLanguage

Termo 11: Idiom
Definição: língua falada em determinado **Place**.

Termo 12: Synonym
Definição: Relaciona conceitos de diferentes línguas que podem ser livremente substituídos (City=Cidade).

Termo 13: Capital
Definição: Relaciona as capitais dos países e estados.

Termo 4: Time[2]
Definição: Essa ontologia fornece um vocabulário para expressar relações topológicas sobre instantes e intervalo de tempo.

[2] http://www.w3.org/TR/owl-time

O corpus refere-se ao Geo-CLEF-2008 que foram coleções de documentos dos jornais Glasgow Harold e Folha de São Paulo. A seguir têm-se alguns exemplos de consultas que foram determinadas pela organização do Geo-CLEF-2008:

(a) Riots in **South American** prisons
(b) **Nobel prize** winners from **Northern European** countries
(c) **Forest fires** on **Spanish islands**
(d) **G7** summits in **Mediterranean countries**
(e) **Bombings** in **Northern Ireland**
(f) Most visited sights in the **capital of France** and its **vicinity**
(g) Unemployment in the **OECD** countries
(h) **Portuguese immigrant communities** in the world
(i) **Natural disasters** in the **Western USA**

Nas consultas (a), (b) e (h) os termos: **South American**, **Spanish islands**, **Mediterranean countries** são instâncias do conceito de Gentílico. Na consulta (b) o termo **Nobel prize** é uma instância do conceito Role. Nas consultas (b), (e) e (i) os termos **Northern European**, **Northern Ireland**, **Western USA** são exemplos de Relações Espaciais do tipo direcionais.

4 Expansão das consultas e criação dos índices

A fase atual da pesquisa está em testar e validar a ontologia desenvolvida. Nesta fase estamos fazendo a expansão semântica das consultas de forma manual. Abaixo tem-se um trecho do arquivo que contem a expansão manual.

```
<parameters>
<query>
<type>Geoclef2008-GH-LATIMES</type>
<number>10.2452/76-GC </number>
<text>
#band(<riot revolt disturbance tumult anarchy uprising>
#or(Argentina Bolivia Brazil Chile Suriname Ecuador
Colombia Guyana Peru Paraguay Uruguay Venezuela French Guiana)
<prison jail penitentiary reformatory>)
</text>
</query>
<query>
<type>Geoclef2008-GH-LATIMES</type>
<number>10.2452/77-GC</number>
<text>
#band (#1(nobel prize) <winner champion master> #or(Denmark Finland
Iceland Norway Sweden Estonia Latvia Belgium Netherlands Luxembourg Ireland
Lithuania United Kingdom Poland Russia))
</text>
</query>
<query>
<type>Geoclef2008-GH-LATIMES</type>
<number>10.2452/78-GC</number>
<text>
```

```
   #band(<#1(sport event) #1(sport happening)#1(sport occurrence) #1(sport
competition)> #or(Algeria Chad Egypt Libya Mali Mauritania Morocco Niger
#1(Western Sahara)Sudan Senegal Tunisia))
   </text>
   </query>
   .....</parameters>
```

Observa-se que na primeira consulta (na letra (a) no sub-topico acima "Riots in South American prisons" utilizou-se o operador Band (operador E lógico) combinado ao operador OR. A expansão dos termos foi feita sob conceito "South American", onde foram adicionados os conceitos: Argentina, Bolívia, Brasil, etc. países que fazem parte da América do Sul, e os conceitos que são sinônimos de "prisons" `<jail penitentiary reformatory>` e sinônimos de "riots"`<revolt disturbance tumult anarchy uprising>`

Será utilizado o programa Lemur toolkit para o teste. O projeto Lemur toolkit é patrocinado pelo ARDA - Advanced Research and Development Activity in Information Technology do National Science Foundation. O Lemur toolkit possui código aberto e tem livre distribuição. O sistema foi concebido para facilitar as pesquisas na área de recuperação de informação e utiliza o Indri Search Engine. Lemur toolkit foi escrito na linguagem C e C++ e atualmente é suportado plataformas Windows (NT e XP) e Unix (GNU/Linux e Unix). Algumas características são: indexa textos em inglês, chinês e árabe, possui técnica de word stemming (Porter e Krovetz), reconhece acrônimos, "stopwords" entre outras. A ferramenta Lemur pode ser obtida no site *http://sourceforge.net/projects/lemur/* e seus respectivos tutoriais. Neste trabalho optou-se pela versão 4.11 para Windows.

O Indri Search Engine é uma combinação do modelo estatístico de linguagem e uma rede de inferência a Figura 4. Utiliza a modelagem estatística de linguagem como estratégia de recuperação de informação. A modelagem lingüística atribui valores probabilísticos, entre 0 e 1, para cada documento significando o escore deste documento. O Indri usa a função logarítmica no calculo dos escores, como o log de zero é igual a um número negativo e log de 1 é igual a zero, conseqüentemente os escores dos documentos são sempre negativos. Os modelos estatísticos de linguagem são essencialmente tabelas de estimativas de probabilidade condicional para algumas ou todas as palavras em uma linguagem que especificam a probabilidade de uma palavra ocorrer na coleção na presença de outras palavras. Por padrão o Indri usa a função a função de probabilidade de DIRICHLET:

```
c(w,D) = contagem das palavras no documento
c(w,C) = contagem das palavras na coleção
|D| = numero de palavras no documento
|C| = numero de palavras na coleção
Numerador = c(w,D) + μ * c(w;C) / |C|
Denominador = |D| + μ
Escore = log( numerador / denominador )
```

Por padrão o valor é de $\mu = 2500$, o que significa que para documentos muito pequenos a diferença entre os escores será muito pequena.

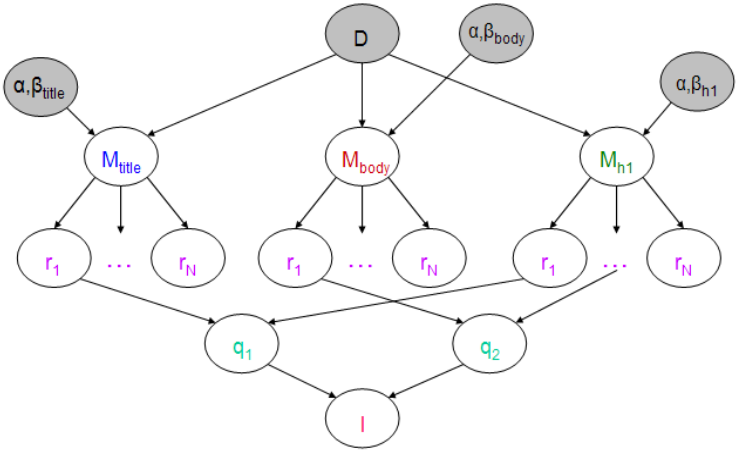

Figura 4. Modelo de recuperação de informação usado pelo Indri

O padrão são consultas com conectivo OR mas se existe a necessidade de que ambos os termos (And) apareçam deve-se usar a expressão #band (xxx) exemplo #band(White house). Esses operadores podem ser combinados para gerar uma consulta mais complexa:

- #combine(dog canine)
- #combine(#1(white house) <#1(president bush) #1(george bush)>)
- #weight(1.0 #1(white house) 2.0 #1(easter egg hunt))

Os principais tipos são:

- #odN(...) -- ordered window -- termos devem aparecer ordenados.
- #N(...) – mesmo que #odN
- #od(...) -- unlimited unordered window – todos os termos devem aparecer ordenados em qualquer posição o texto.
- #uwN(...) unordered window – todos os termos devem aparecer dentro de um tamanho especificado em qualquer ordem.
- #uw(...) -- unlimited unordered window – todos os termos devem aparecer no texto em qualquer ordem.
- #combine
- #weight
- #not
- #max
- #or
- #band (boolean and)
- #wsum

A próxima fase da pesquisa é a preparação do corpus e do índice. Foi necessário um parser que convertesse os documentos que estavam no formato SGML para o formato TrecText que é nativo do Lemur. Abaixo o formato do arquivo do Corpus em TrecText.

```
<DOC>
<DOCNO> document_number </DOCNO>
<TEXT>
Index this document text.
</TEXT>
</DOC>
```

O Lemur aceita dois tipos de índices KeyfileIncIndex e IndriIndex (Tabela 1). O IndriIndex permite apenas a recuperação de consultas através da IndriQuery Language, enquanto que a Inquery Query Language é aceita pelos dois indexadores (KeyfileIncIndex e IndriIndex). O IndriIndex pode ser gerado tanto pelo aplicativo IndriBuildIndex quanto pelo BuildIndex e o índice Keyfile pode ser gerado tanto pelo BuildIndex quanto pelo RetVal (Tabela 2).

Tabela 1. Sumario das características de cada um dos tipos de índice do Lemur

Index Name	Extension	File Limit	Stores positions?	Loads fast?	Disk space usage	Applications	Add documents to Index*
KeyfileIncIndex	.key	no	yes	yes	Average	BuildIndex	yes, use BuildIndex
IndriIndex		no	yes	yes	most (automatically stores compressed version of original documents)	BuildIndex or IndriBuildIndex	yes

* Supports adding new documents to index, not updating existing documents.

Tabela 2. Tipos aplicativos para geração de índices presentes no Lemur

Applications

Feature	Keyfile Index	Indri Index
Index Building	BuildIndex	IndriBuildIndex or BuildIndex
Batch Retrieval	RetEval	IndriRunQuery or RetEval

Os parâmetros para a criação dos índices que são suportados pelo IndriBuildIndex são delimitados por tags <parameters> </parameters>. Pode ser definido mais de um parâmetro de acordo com a necessidade. O modelo padrão de um arquivo de parâmetros tem a seguinte configuração é conforme mostrado abaixo, mas como utilizamos a ferramenta gráfica não foi necessário a criação do arquivo de parâmetros. Onde:

`<index>/home/lemur/testindex</index>` Mostra o caminho onde será criado seu arquivo de indice.

`<corpus>`

```
    <path>/home/lemur/testdata/firstCorpus</path>
    <class>trectext</class>
</corpus>
```

É o caminho onde se encontra o **corpus** sobre o qual serão feitas as consultas. Pode-se ainda especificar os seguintes itens:

- **Class**: define o tipo de arquivo do corpus no arquivo de parâmetros ou pela linha de comando: -corpus.class=trecweb (tipo de arquivo usado neste trabalho)
- **Annotations**: define o caminho para o arquivo que contém a anotação caso exista sobre o corpus. Quando se quer indexar também as tags da linguagem HTML cria-se este arquivo ou de corpus que contenham conteúdo PHP .
- **Metadata**: caminho para o arquivo de metadados, ou em linha de comando: -corpus.metadata=/path/to/file..

```
<stemmer><name>krovetz</name></stemmer>
<stopper><word>stopword</word></stopper>
```

- **Stopper**: contém o arquivo de stopwords, o Lemur já disponibiliza uma lista de stopword para o ingles. Em linha de commando: -stopper.word=stopword.
- **Stemmer**: especifica o modelo de Stemmer, Lemur disponibiliza dois tipos de Stemmer: 'Porter' or 'Krovetz' (case insensitive). Na linha de comendo: -stemmer.name=stemmername.

5 Discussão

A pesquisa avança no sentido de validar o uso de ontologia para dar suporte semântico documentos a serem recuperados e a ferramenta de busca, de modo que seja possível obter resultados da pesquisa que possam refletir de forma mais adequada a intenção e as conceitualizações do usuário. Os próximos passos são em primeiro lugar fazer as simulações usando a ferramenta Lemur, e em segundo avaliar os resultados em função destes resultados e verificar a melhoria do desempenho que relações semânticas possam ter em sistemas de recuperação de informação.

Referências

[1] KLINGBERG, T. (2008) **The overflowing brain: information overload and the limits of working memory.** Disponivel em < http://books.google.com.my/books?q=information+overload> Acesso em março 2010.

[2] BERNERS-LEE T.; Hendler, J. and Lassila, O. The Semantic Web. Scientific American, may 2001, pp. 28-37.

[3] WORLD WIDE WEB CONSORTIUM – W3C, acesso em 29 nov. 2009, <http://www.w3.org/>.

[4] FENSEL, D. et al, On-To-Knowledge: Semantic Web Enabled Knowledge Management, Disponível em <http//:*www.few.vu.nl/~frankh/postscript/WI-book03.pdf*>, *Acesso em dez 2009.*

[5] FENSEL, D. Ontologies: Dynamic Networks of Formally Represented Meaning. **Semantic Web Portal Project**. Disponível em < http://sw-portal.deri.at/papers/publications/network.pdf >, Acesso em 21 Mai. 2008.

[6] GUARINO, N. e GIARETTA, P. Ontologies and Knowledge Bases. Towards a Terminological Clarification. Padova, Italy, 1995. Disponível em <http://www.loacnr.it/Papers/KBKS95.pdf>, Acesso em 13 out 2008.

[7] GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. International journal of Human-Computer Studies, 1995, no. 43, p. 907–928.

[8] KASHYAP, V. e SHETH, A. Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In Papazoglou, M. P. and Schlageter, G., editors, Cooperative Information Systems, 1997, p.139–178. Academic Press, San Diego.

[9] GUARINO, N. Formal ontology and information systems. In: N. Guarino, editor, *Formal.* Amsterdam, The Netherlands . Publisher IOS Press. 1998 p. 3-15.

[10] GUARINO, N. e WELTY, C. 2000. A formal ontology of properties. In Dieng, R., ed., Proc. of EKAW'00, LNCS. Springer Verlag.

[11] MAEDCHE, A. e STAAB, S. 2001, "Ontology learning for the Semantic Web." *IEEE Intelligent Systems*, vol. 16, no. 2, pp. 72-79.

[12] TOMAI, E . e KAVOURAS, M. 2004, "From "Onto-GeoNoesis" to "Onto-Genesis": The Design of Geographic Ontologies." *GeoInformatica*, vol. 8, no. 3, pp. 285-302.

[13] GUIZZARDI, G. 2005. Ontological Foundations for Structural Conceptual Models. Phd thesis, University of Twente, The Netherlands. Telematica Instituut Fundamental Research. Series No. 15.

[14] NOY, N. F. e McGUINNESS, D. L. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University: Stanford. (2005). Disponível em <http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html >, Acesso em 13 out 2008.

[15] GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ,M. e CORCHO O. Ontological engineering : with examples from the areas of knowledge management, e-commerce and the semantic web, 2004.

[16] DAHLBERG, I. Teoria do conceito. Revista Ciência da Informação, v.7, no.2, p.101-107. 1978.

[17] RANGANATHAN, S R. 1985. Faceted analysis. En Chan, L. M. et al., eds.Theory of subject analysis. Littleton, CO: Libraries Unlimited. p. 86-93.