SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
3ºONTOBRAS
30 e 31 de agosto de 2010 - Florianópolis/SC

MOFI: Um Modelo para Recuperação de Informação baseado em Ontologias, Folksonomia e Indexação Automática de Conteúdo

Uedson Santos Reis (SENAI) uedsonreis@gmail.com
Eduardo Manuel de Freitas Jorge (UNEB) emjorge1974@gmail.com
Hernane Borges Barros Pereira (SENAI) hernanebbpereira@gmail.com

A maioria dos motores de busca da web leva a sintaxe, e não a semântica, em consideração no momento de realizar uma busca, o que nem sempre traz um resultado satisfatório. Este artigo propõe a construção do modelo MOFI que apoia a construção de bases de dados semi-estruturados através da associação de técnicas como Ontologias, Folksonomia e Recuperação de Informação, como a indexação automática de arquivos. E ainda apresenta e avalia quatro tipos de busca que podem ser feitos com base nas relações de uma Ontologia. Este modelo beneficia a difusão do conhecimento, uma vez que os motores de busca poderão entender melhor a semântica de cada dado, obtendo um retorno mais preciso.

Palavras-chave: Ontologia, Folksonomia, Indexação e Recuperação de Informação

1. Introdução

Pessoas ao redor do mundo vêm alimentando bases de conhecimento na *web*, principalmente através de serviços como sites de relacionamento, redes sociais, enciclopédias virtuais, *blogs* e fóruns de discussão. Essas ferramentas facilitam a publicação de conteúdo por parte do usuário, aumentando a quantidade de dados armazenados nas diversas páginas *web*, potencializando a capacidade do usuário de difundir conteúdo na *web*.

A maior parte desse conteúdo só tem significado para os seres humanos, não podendo ser entendido, em um contexto semântico, pelas máquinas. Desta forma, torna-se difícil a tarefa de transformação desse conteúdo em conhecimento, facilitando o seu acesso e a sua utilização por parte do usuário. Como reflexo disso a maioria dos motores de busca na *web* acaba levando a sintaxe, e não a semântica, em consideração no momento de realizar uma busca, o que nem sempre traz um resultado satisfatório (BREITMAN, 2005). A realização de tarefas por parte de agentes lógicos na *web* também fica comprometida também pelo mesmo motivo (GRUBER, 1993).

Estruturas que representem o conhecimento, como por exemplo, Ontologias, podem auxiliar a atribuir mais semântica aos conteúdos. Ontologias, no âmbito da Ciência da Computação, são estruturas formais de conceitos e seus relacionamentos que especificam a conceitualização de um determinado domínio do conhecimento, por isso podem ser manipuladas e entendidas por sistemas computacionais (MOREIRA, et al., 2004).

A utilização da Folksonomia também pode auxiliar na ampliação da semântica desses dados. Essa técnica já é bem popular na *web*, nela o próprio usuário classifica seu conteúdo, atribuindo uma ou mais tags ao mesmo, isso facilita a recuperação desse conteúdo por intermédio de mecanismos de busca baseados nessa técnica (PRIMO, 2006). Entende-se neste trabalho, tag como etiqueta, rótulo ou palavra para fins de classificação de conteúdo. E segundo pesquisa feita, por este autor em conjunto com outros pesquisadores, a Folksonomia também pode ser utilizada de forma integrada com as Ontologias. Desta forma, além de se beneficiar do formalismo da Ontologia para classificação dos dados e permitir a manipulação por parte de softwares, também agregar os benefícios do conhecimento coletivo obtidos com a Folksonomia (REIS, et al., 2009).

Portanto, o problema que será estudado é como se utilizar dessas novas tecnologias para aperfeiçoar as técnicas convencionais de Recuperação de Informação. Ressalta-se que apesar da evolução da técnica de indexação de arquivos, esta ainda não fornece a melhor solução para localização de conteúdos na *web*. Por usar as palavras chave como base, o resultado da busca quase sempre traz uma quantidade muito grande de dados, que em sua maioria não tem a informação desejada. Também existe o problema clássico relacionado com a semântica, já que esses motores de busca só entendem a sintaxe de uma palavra chave. Por exemplo, o retorno de uma busca pela palavra *artigo* trará resultados ligados a artigos acadêmicos e a livros de português com definições e exemplos sobre artigos definidos e indefinidos.

O objetivo do artigo consiste em apresentar um modelo, denominado de MOFI, que se utilize de Ontologias, Folksonomia e Técnicas de Indexação Automática de Arquivos para aperfeiçoar os processos de armazenamento e recuperação de dados semi-estruturados (arquivos, textos ou fotos).

O MOFI é um modelo computacional que se baseia nos seguintes elementos: (i) Um módulo que utilize um motor de busca baseado em indexação automática de arquivos para realização de busca sintática, como feito em sites de busca como *Google*, *Yahoo* e *AltaVista*; (ii) Um módulo que adote a técnica de Folksonomia para a indexação manual dos arquivos, permitindo ao usuário classificar o conteúdo com termos inseridos livremente ou com os conceitos da Ontologia, seguindo a técnica sugerida por REIS, et al, (2009); (iii) Um módulo que gerencie consultas a uma Ontologia a fim de aperfeiçoar a busca, contextualizando-a ou tentando localizar termos sinônimos; (iv) Um sistema colaborativo baseado na web para armazenamento e recuperação de dados semi-estruturados que integre todos esses componentes.

O MOFI tem como finalidade melhorar o compartilhamento e otimizar a recuperação do dados disponíveis na web, assim como permitir às máquinas, por intermédio de agentes lógicos, entender e manipular melhor esse conteúdo facilitando tarefas desenvolvidas pelos seres humanos. A difusão do conhecimento também será facilitada nesse contexto, uma vez que os motores de busca poderão entender melhor a semântica de cada dado manipulado no momento da busca, obtendo um retorno mais preciso.

Este artigo está estruturado da seguinte forma: a seção 2 apresenta a técnica de Indexação de arquivos como base para o componente de indexação automática; na seção 3 a Folksonomia será definida e detalhada para utilização no componente de indexação manual; a seção 4 explica como o motor de busca funciona manipulando ontologias antes de realizar as buscas; já na seção 5 a integração dos componentes é mostrada e ilustrada; por fim, são apresentadas as considerações finais.

2. Recuperação de Informação: Método Convencional

A estrutura da *web* nem sempre facilita a recuperação ou localização de um conteúdo disponibilizado nela, já que ficam armazenados geralmente em uma estrutura centralizada de hierarquia de pastas e subpastas. Para localizar um determinado conteúdo o usuário tem que saber o endereço do site e qual o caminho exato do conteúdo dentro do site, o que na maioria das vezes não é cômodo.

Alguns pesquisadores e colaboradores da rede mundial de computadores estão tentando facilitar o acesso a essa grande quantidade de informação que está disponível e descentralizada na *web*. Estudos e técnicas como a Inteligência Artificial, a Web Semântica, e a Indexação de Arquivos junto com Algoritmos de Busca, estão sendo desenvolvidos com esse objetivo (GOSPODNETIC, et al., 2005).

Dentro desse contexto, surgiram os motores de busca (ex.: Google, Yahoo), desenvolvidos para facilitar a localização de conteúdo em toda a *web*. Esses motores extraem os termos mais freqüentes dentro de um texto, denominados de palavras-chave, para indexar e buscar arquivos e páginas *web*, facilitando a vida de estudantes, pesquisadores e outros usuários em todo o mundo.

Para localizar arquivos indexados, basta que os motores de busca leiam os índices dos arquivos e retornem os documentos que tiverem em seus índices as palavras-chave da busca. Geralmente a ordenação desse resultado leva em consideração o número de vezes que as

palavras-chave aparecem juntas no arquivo, quanto mais isso ocorre mais relevante é o arquivo. Uma das ferramentas mais utilizadas para auxiliar desenvolvedores a inserir indexação e busca em suas aplicações é o Apache Lucene (GOSPODNETIC, et al., 2005).

O Lucene é uma biblioteca de Recuperação de Informação de alto desempenho. Ele é um projeto livre de código aberto, escrito em Java e mantido pela fundação Apache. O Lucene utiliza técnicas de indexação de arquivos para aperfeiçoar suas buscas, que são realizadas em cima desses índices gerados. Esse tipo de técnica é utilizado na ferramenta de busca do Google, uma das ferramentas de busca mais utilizadas na *web* atualmente (GOSPODNETIC, et al., 2005).

O Lucene indexa os dados armazenados pela aplicação, que podem estar em bases de dados distintas, gerando um índice onde serão feitas as buscas do usuário. A gerência dos dados e a visualização do conteúdo retornado para o usuário ficam a cargo da aplicação que estiver utilizando o Lucene.

3. Recuperação de Informação com Folksonomia

Alguns serviços na web já estão tentando aperfeiçoar o seu processo de armazenamento e recuperação. Sites de compartilhamento de conteúdo, como *Flickr* e *Del.icio.us*, e *webmails*, como *Gmail*, permitem que seus usuários aumentem a semântica de seus conteúdos através das *tags* (ou rótulos para fins de classificação). O usuário atribui uma ou mais *tags* ao conteúdo, que são anexadas ao mesmo através de um metadado ou num banco de dados. Isso facilita a recuperação desse conteúdo por intermédio de mecanismos de busca que recuperam o conteúdo a partir dessas *tags*, essa estrutura é denominada de Folksonomia (VANDER WAL, 2007).

A palavra Folksonomia tem origem nas palavras *folk* e *taxonomia*. Onde *folk* significa povo e *taxonomia* é o estudo da classificação das coisas. A idéia, por traz da Folksonomia, é que o usuário classifique seus dados de forma livre, viabilizando uma melhor eficiência no armazenamento e na recuperação desses dados ante a ação de motores de busca (VANDER WAL, 2007).

No contexto da Folksonomia as *tags* são atribuídas pelo usuário, que tem total liberdade no momento de classificar seu conteúdo. Entretanto, além dos erros de sintaxe, cada pessoa classifica seus conteúdos de uma maneira própria, seguindo conceitos que para ela tem sentido ou significado, isso facilita a recuperação desse conteúdo por parte da mesma pessoa

que o classificou, porém a mesma facilidade nem sempre é encontrada se outra pessoa, que não participou da classificação, tentar recuperar esse conteúdo. Esse ponto pode ser considerado uma desvantagem da Folksonomia quando utilizada sozinha para o armazenamento e recuperação de conteúdo (REIS, et al., 2009).

Um aspecto que pode ser considerado positivo na Folksonomia é a capacidade de cruzamento das próprias *tags*, isto facilita a extração de informações armazenadas implicitamente nos dados rotulados. Um exemplo bem comum é quando um documento recebe a mesma *tag* varias vezes de pessoas diferentes, isso na maioria dos casos pode ser considerada uma classificação conceitualmente apropriada, uma vez que o conteúdo tem a mesma representatividade para vários outros usuários. Da mesma forma quando uma *tag* é atribuída a um documento por apenas dois ou três usuários isso provavelmente será considerada uma classificação específica dessas pessoas (REIS, et al., 2009).

O modelo computacional proposto neste artigo, também proverá o esquema da Folksonomia, que terá suas tags vinculadas aos arquivos por intermédio de um banco de dados. Esse banco de dados irá manter informações dos arquivos, como endereço onde está armezenado, uma breve descrição de seu conteúdo, e principalmente as *tags* relacionadas a ele. Este modelo também permitirá ao usuário utilizar termos vindos da Ontologia, seus conceitos ou classes, como *tags*, classifincando ou buscando conteúdo.

O uso da folksonomia tem um ponto crítico neste modelo, pois requer um espaço de armazenamento muito extenso para as *tags*. Por ser uma classificação feita de forma livre, o usuário pode atribuir quantos termos desejar o que, no longo prazo, pode aumentar muito o número de registros no banco de dados. Por esse motivo, recomenda-se a exclusão no banco de dados das tags que não estiverem sendo utilizadas a cada período determinado de tempo.

4. Ontologia

Existem varias definições para o termo Ontologia podem variar a depender do contexto utilizado. Essas definições podem ser distintas ou complementares, entre si, para um mesmo conceito ou domínio, e o seu significado tende a variar de acordo com o objetivo de seu uso (BREITMAN, 2005).

Thomas Gruber (1993) definiu Ontologia como estrutura de conceitos e seus relacionamentos que especifica a conceitualização compartilhada de um determinado domínio em nível semântico. Em outras palavras, uma Ontologia define formalmente conceitos e restrições de

um determinado domínio, através de uma estrutura de relacionamentos. Isso permite que a ontologia tanto seja manipulada pela máquina como também absorva conhecimento consensual, podendo ser entendida também pelos seres humanos (BREITMAN, 2005).

Na ciência da computação o termo ontologia começou a ser utilizado na área da Inteligência Artificial, em projetos para formação de grandes bases de conhecimento (MOREIRA, et al., 2004). Ela tem a estrutura de uma rede semântica, porém suas regras de relacionamento são mais rigorosas.

Alexander Maedche, segundo Breitman (2005), descreveu os cinco elementos como sendo componentes básicos de uma ontologia, são eles: (i) Conceitos (classes): definem as sub-áreas de um domínio de interesse; (ii) Relacionamento: Formas de interligar os conceitos. Um possível relacionamento, por exemplo, entre homem e criação é *criador_de;* (iii) Hierarquia: formas como os conceitos se organizam dentro de um domínio. Como uma taxonomia, onde o carro é um sub-conceito de veiculo; (iv) Funções: determinam formas especiais de relacionamento, onde um determinado número de elementos se relaciona com apenas um elemento. Por exemplo, uma função *propriedade_de* pode interligar um carro e uma moto a apenas uma pessoa; (v) Axiomas: determinam verdades sobre um determinado domínio. Como por exemplo, para ser mãe, uma mulher, tem de ter no mínimo um filho ou filha.

Através desses componentes, pode-se utilizar Ontologias também para aperfeiçor processos de armazenamento e recuperação de conteúdos. Um exemplo que pode ser citado é o das organizações Globo, com um projeto que visa estruturar o seu portal através do uso de Ontologia. A ideia é utilizar uma Ontologia que trate de assuntos gerais, como pessoas ou localidades, e mais quatro ou cinco Ontologias especializadas, abordando domínios como Esporte, Política ou Economia. Através dessas Ontologias as buscas feitas no portal *Globo.com* poderão ser mais específicas.

Usando o poder das relações entre conceitos de uma Ontologia o portal pode realizar uma busca por todos os gols perdidos por um determinado jogador e retornar um resultado satisfatório. Ao invés de realizar uma busca sintática, como feita normalmente pelos motores de busca, com os termos gols perdidos de Paulo, onde o resultado, certamente, conteria alguns gols feitos por Paulo e não os perdidos. Mesmo a solução de associar tags (palavras-chave) as informações do site (encontrada em alguns serviços na *web*) pode contribuir para especializar buscas, porém esta abordagem é limitada, pois as tags não se relacionam. Isto é, mesmo que as tags gols, perdidos e Paulo estejam vinculadas ao vídeo, a busca seria feita sem considerar

uma relação entre as tags. Já com uma Ontologia de futebol que pode possuir duas relações entre jogador e gol, como as propriedades gols perdidos e gols feitos a busca seria feita com mais facilidade, considerando essa relação semântica.

5. Modelo MOFI

Às técnicas de Ontologia, Folksonomia e Indexação Automática de arquivos foram associadas para formar o modelo MOFI, visando estruturar bases dados semi-estruturados (arquivos, textos ou fotos). Uma das características do modelo MOFI é na realização de uma busca consultar uma Ontologia, escolhida previamente pelo usuário, para aprimorar a busca. Outro ponto é a utilização dos conceitos, ou classes, de uma Ontologia como *tags* para classificar, por meio do componente de indexação manual, ou buscar os conteúdos, ai através de ambos os componentes. O usuário poderia escolher esses conceitos através de uma metafora visual que disponibilizaria esses conceitos como uma nuvem de tags (REIS, et al., 2009).

O motor de busca do MOFI possui dois módulos para execução da busca. O primeiro é o módulo de indexação automática, que possui o índice dos termos encontrados nos conteúdos armazenados, por meio desse módulo pode-se localizar textos existentes nos conteúdos dos arquivos. O segundo é o módulo de indexação manual, que tem seu funcionamento baseado na tecnica da Folksonomia, o que permite ao usuário localizar seus conteúdos através de tags, ou rótulos, atribuídas aos conteúdos previamente.

O motor de busca irá tentar localizar na Ontologia alguma das palavras-chave informadas pelo usuário, caso encontre alguma delas, montará as chaves de busca de forma a implementar buscas baseadas nas ligações do termo encontrado com outros termos da Ontologia, os termos que não forem encontrados serão includos na chave de busca diretamente sem refinamento.

O diferencial do MOFI está nos seus 4(quatro) os tipos de processamento de busca. Detalha-se cada tipo a seguir através de um exemplo ilustrativo no contexto de futebol: (i) Busca por sinônimos – caso alguma das palavras-chave, que foram encontrados na Ontologia, tenham uma relação de equivalência com algum outro termo esses serão incluídos na chave de busca com o operador lógico *OR*, de forma a também ser procurado, um exemplo dessa chave seria *Aluno OR Discente*. O objetivo desse tipo de busca é localizar termos sinônimos não encontrados em busca baseadas em estrutura sintática; (ii) Busca por hierarquia - caso alguma das palavras-chave, que foram encontrados na Ontologia, tenham uma relação de hierarquia, sendo ela a superclasse, com algum outro termo esses serão incluído na chave de busca com o

operador lógico *OR*, de forma a também ser procurado, um exemplo dessa chave seria *Jogador OR Goleiro OR Jogador de Linha* sendo *Jogador* o termo buscado. O objetivo desse tipo de busca é encontrar conceitos que não são sinônimos, mas que são um tipo do conceito procurado;(iii) Busca contextualizada com o termo buscado - caso alguma das palavras-chave, que foram encontrados na Ontologia, tenham outros tipos de relação, exceto equivalencia e hierarquia para baixo, com algum outro termo esses serão incluídos na chave de busca com os operadores lógicos *AND* e *OR*, um exemplo dessa chave seria *Jogador AND (Futebol OR Time OR impedimento)* sendo *Jogador* o termo busca. O objetivo desse tipo de busca é filtrar os resultados da busca, focando no contexto definido pela Ontologia escolhida. (iv) Busca contextualizada sem o termo buscado - caso alguma das palavras-chave, que foram encontrados na Ontologia, tenham outros tipos de relação, exceto equivalencia e hierarquia para baixo, com algum outro termo esses formaram sozinhos a chave de busca com o operador lógico *AND*, um exemplo dessa chave seria *Jogador AND Time AND impedimento AND Gol)* sendo *Futebol* o termo buscado. O objetivo desse tipo de busca é retornar conteúdos que tenham a há ver com o contexto definido pela Ontologia escolhida mesmo que o termo buscado não esteja presente, já que é possível que um texto fale de futebol sem mencionar a palavra futebol.

Com os termos relacionados encontrados na Ontologia de Futebol, esse motor de busca monta a chave de busca de forma a realizar consultas que tragam todos os arquivos que contiverem a palavra futebol (termo informado pelo usuário) e algum desses outros termos (informados pela Ontologia como termos ligados diretamente a futebol). E também os arquivos que não tenham a palavra futebol, mas que tenham mais de 80%, por exemplo, desses outros termos (os informados pela Ontologia).

Um dos diferenciais do modelo MOFI, em relação a buscas sintáticas, é que existe uma diminuição do retorno de conteúdos sobre futebol americano, aumentando o retorno de conteúdos que falem sobre o futebol tradicional, por exemplo. E ainda permite ao motor de busca encontrar textos que falem de futebol sem ter a palavra futebol, especificamente, como um texto que fale de um gol feito pelo jogador A do time B em impedimento aos 45 minutos do segundo tempo.

Utilizando a propriedade de equivalência, existente na Ontologia, ainda podem-se obter termos equivalentes ou sinônimos de outros termos. Desta forma, a Ontologia exemplo de domínio Futebol terá uma definição de equivalência do conceito Arbitro para o conceito Juiz, o que permitirá ao motor de busca proposto realizar a busca pelo termo sinônimo de Arbitro.

O que em um motor de busca apenas sintático, isso não seria possível. A busca por hierarquia de conceitos é realizada quando a palavra-chave que se deseja buscar está no contexto da Ontologia e tem conceitos abaixo em sua hierarquia. Por exemplo, as palavras Jogador e Goleiro. Elas não significam a mesma coisa, não são sinônimos, mas o Goleiro é um tipo de Jogador, portanto se procuramos por um Jogador esse Jogador pode ser um Goleiro, o que não ocorre no caso contrário, quando procuro um Goleiro não posso retornar qualquer Jogador, tem que ser um Goleiro e não um Jogador de Linha por exemplo.

Caso o motor de busca proposto não encontre nenhuma das palavra-chave, solicitadas pelo usuário, na Ontologia ou o termo encontrado não possua ligações que permitam realizar os quatro tipos de busca, as chaves de busca serão montadas de forma convencional, onde os termos solicitados são pesquisados sem a inclusão de novos termos ou operadores lógicos.

5.1 Validação do Modelo MOFI

Para validar MOFI foram feitos alguns testes de consulta a uma base de arquivos de tipos diversos, *pdf, doc,* imagens e escritos em língua portuguesa e inglesa. Foram feitas buscas através do Lucene e através de um algoritmo implementado para simular o funcionamento do modelo MOFI. Em todos os casos o resultado foi satisfatório. Nas buscas contextualizadas o algoritmo conseguiu filtrar melhor os resultados, numa busca feita pela palavra *time* o Lucene trouxe muitos artigos em inglês que tinham a palavra *time* em inglês, que significa tempo em português, e trouxe também três textos que falavam realmente de times de futebol. Já com o algoritmo, obteve-se o mesmo resultado quando uma Ontologia não era selecionada, e quando escolhemos uma Ontologia de Futebol para aperfeiçoar a busca, o resultado foi mais enxuto, cerca de 10% dos arquivos que tinham a palavra *time* em inglês foram retornados, mais os três textos que falavam de times de futebol.

Também se obteve sucesso quando foram feitas as buscas por sinônimos, que estavam modelados na Ontologia, e por hierarquia, utilizando os exemplos citados acima de *Árbitro* e *Juiz*, e *Jogador* e *Goleiro*. Assim como o segundo tipo de busca contextualizada, onde o teste foi feito utilizando um dos três arquivos que falavam de times de futebol. Esse arquivo em específico, fala da reclamação de um dirigente quanto a um gol feito, em impedimento, contra o seu time. Ao realizar uma busca pela palavra futebol, esse texto que não tinha a palavra futebol foi retornado no resultado da busca, pois continha termos que estavam ligados diretamente a futebol permitindo o retorno acertado do arquivo.

6. Considerações Finais

Neste artigo foi apresentado a proposta de um modelo MOFI que tem como principal caracteristica a utilização conjunta de tecnicas de Folksonomia e Indexação de Automática Arquivos, e mais a utilização das relações de uma Ontologia para aperfeiçoamento das busca, realizando buscas por sinônimos e contextualizadas.

Um dos diferenciais do modelo MOFI são os seu 4 (quatro) tipos de processamento de busca poutados em semânica apresentados na seção 5. Como trabalhos futuros sugere-se pesquisas qualitativas e quantitativas com o intuito comparativo de desempenho, usabilidade e aceitação do modelo proposto em relação aos demais métodos de Recuperação de Informação. Propõe-se também a evolução da ferramenta, possibilitando a a utilização dos axiomas para também aperfeiçoar as buscas.

Bibliografia

BREITMAN Karin. Web Semântica a Internet do Futuro [Livro]. - Rio de Janeiro : LTC, 2005.

GOSPODNETIC Otis e HATCHER Erik Lucene In Action [Livro]. - [s.l.] : MANNING, 2005. - Vol. 74.

GRUBER Tom. A Translation Approach to Portable Ontology Specifications. - Abril de 1993.

MOREIRA Alexandra, ALVARENGA Lídia e OLIVEIRA Alcione. O Nível do Conhecimento e os Instrumentos de Representação: Tesauros e Ontologias. - Dezembro de 2004.

PRIMO Alex O Aspecto Relacional das Interações na Web 2.0 [Conferência] // XXIX INTERCOM: Congresso Brasileiro de Ciências da Comunicação. - Brasília : [s.n.], 2006.

REIS Uedson [et al.] Ontology Tagging - Um componente para Integração de Ontologia com Folksonomia [Conferência] // ERBASE - Escola Regional Bahia Sergipe. - Salvador : [s.n.], 2009.

VANDER WAL Thomas. Folksonomy Coinage and Definition [Online] // Vanderwal.net. - Fevereiro de 2007. - Agosto de 2008. - http://vanderwal.net/folksonomy.html.