UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Informática - PPGI

Mapeamento Semântico entre Ontologias utilizando Axiomas e Classificação

Fabrício D'Morison da Silva Marinho

Orientadora: Virgínia Brilhante

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Ciências Exatas da Universidade Federal do Amazonas como requisito parcial para obtenção do título de Mestre em Informática.

MANAUS – AM

MARÇO DE 2007

Mapeamento Semântico entre Ontologias Utilizando Axiomas e Classificação

Fabrício D'Morison

Universidade Federal do Amazonas (UFAM)
Departamento de Ciência da Computação
Manaus, AM, Brasil

fabricio.dmorison@gmail.com

Resumo. *Descobrir automaticamente relações semânticas entre ontologias é uma tarefa de grande importância, visto que isto pode ser usado para que sistemas distintos se comuniquem em nível semântico, ou seja, interoperar com base em conhecimento. Este trabalho propõe um estudo de como o uso de classificação automática e de axiomas de domínio pode evidenciar similaridade semântica, potencializando a precisão de mapeadores de ontologias (os quais até hoje são fortemente baseados apenas em evidências léxicas, sintáticas e na hierarquia de conceitos e relações), derivando métricas mais apropriadas tanto para descobrir mapeamentos quanto para qualificar cada mapeamento, associando-o com a relação semântica mais adequada, gerando assim axiomas de mapeamento. O foco principal está em, combinar estratégias de mapeamento iniciais (gerados por classificação, por exemplo) com estratégias baseadas em axiomas de domínio, pois acreditamos que, dada a árvore de dependência entre eles, quanto menos dependente for um axioma mais fácil será compará-lo com outros axiomas e mais freqüente será encontrá-lo em ontologias distintas.*

1. INTRODUÇÃO

1.1. MOTIVAÇÃO E OBJETIVOS

Um mundo em que computadores sejam capazes de discernir e aprender é o sonho de muitos, dado que a automatização de atividades humanas é a grande utilidade da informática, a qual trabalha e desenvolve tecnologias para si mesma e para todas as outras áreas de conhecimento. Muitos processos necessitam não mais que uma receita exata com instruções finitas e explícitas de como processar uma atividade e por isso são facilmente automatizáveis. Tal qual são sistemas de aeroportos, de locadoras, de livrarias *on-line*, de *workflow*, etc. No entanto, a automatização de atividades que demandam passos subjetivos ainda é um desafio à Ciência da Computação (CC), que busca através da Recuperação de Informação (RI), Inteligência Artificial (IA) e Aprendizado de Máquina, meios automáticos para tomada de decisões subjetivas, situação na qual mesmo uma pessoa pode se deparar com dúvida.

Descrever conhecimento através de ontologias vem se tornando cada vez mais popular, situação esta motivada em especial pela Web Semântica, bem como por outras aplicações que necessitam representar, além de dados, informações formalmente estruturadas. Inevitavelmente, é comum que estas aplicações acabem usando representações múltiplas para as mesmas informações, implicando na necessidade de integração das informações. Particularmente, as características ímpares da Web, como o crescimento exponencial do seu já imenso volume de dados, a diversidade de assuntos e a grande quantidade de usuários, levarão a um desenvolvimento distribuído de inúmeras ontologias para a Web Semântica, decorrendo em informações multiplamente representadas, tornando a integração das ontologias uma questão crucial, o que seria muito trabalhoso e demandaria muito tempo fazer manualmente. De fato, os recursos da Web são imbuídos de identidade ou relevância semântica e representam um novo desafio às pesquisas que envolvem busca de informação e extração de conhecimento. Esta situação também é comum em vários outros sistemas que lidam com muito conteúdo, especialmente os de Data Warehouse e Tecnologia da Informação (TI). A este problema de integração ontológica chamamos de *mapeamento entre ontologias.*

A estruturação de dados e informações através de formalização aumenta a homogeneidade destes, facilitando a atuação de sistemas como agentes inteligentes e máquinas de busca, as quais até hoje ignoram informação semântica. Mas se por um lado, ontologias são uma boa estratégia de homogeneização, por si só elas inserem outro tipo de heterogeneidade em termos de semântica, dificultando os mapeamentos [4]: em ontologias que descrevem o mesmo conhecimento, um mesmo conceito pode receber rótulos diferentes ou ser definido de maneiras diferentes, ou ainda, pode ocorrer que conceitos com o mesmo rótulo tenham significado diferente.

No estado da arte, a dificuldade é mais do que automatizar a tarefa de mapeamento, diminuindo a interação humana nas estratégias semi-automáticas, mas principalmente criar as heurísticas que desempenhem satisfatoriamente a tarefa. Este problema é muito mais difícil no âmbito de *mapeamentos semânticos*: uma abordagem mais ambiciosa e necessária é levar em consideração como os axiomas são mapeados, mas a literatura é pobre em métricas e processos que usam os axiomas; até a três anos atrás, não existiam trabalhos comparando as interpretações das entidades ontológicas [5,10], os quais concordam que o único algoritmo relevante até então era [8] e até hoje os trabalhos nesta área ainda são pioneiros. Na prática, quase todos os trabalhos recaem na utilização de apenas algumas poucas evidências léxicas, sintáticas e baseadas em relações de herança para calcular um coeficiente de similaridade entre os elementos das ontologias envolvidas, indicando como equivalentes os elementos que parecerem ser mais similares. Esta forma de construir mapeamentos não garante respeitar e manter a estrutura lógica das ontologias envolvidas, ou seja, não garante a manutenção da semântica por permitir a criação de mapeamentos que criarão inconsistências durante a interoperação entre as ontologias envolvidas e isso, obviamente, não é desejável em nenhuma aplicação. Em suma, a explicação dada pelos autores é a seguinte [7]: *"Certas evidências, como as propriedades algébricas, equivalências e disjunções, não são suficientemente usadas pela comunidade ao desenvolver ontologias para serem consideradas como material de similaridade. Já no caso dos axiomas, que incluem as regras e as restrições, não existe pesquisa nem suporte prático suficiente"*.

Sendo assim, a falta de satisfatória abordagem semântica sobre mapeamento entre ontologias motivou o trabalho desta proposta, que objetiva uma investigação empírica acerca de métodos, métricas, evidências e ferramentas para estabelecer um *operador semântico* de boa qualidade para comparação entre ontologias. Para a criação de tal operador, será dada ênfase à utilização de classificadores e à utilização de axiomas de domínio (os quais são imbuídos de semântica), que são interessantes ao trabalho e justificam nossos fins.

1.2. CONCEITOS E TERMINOLOGIAS

1.2.1. ONTOLOGIAS

Na Filosofia, ontologias dizem respeito ao estudo do que existe no mundo, ao estudo do "ser". Porém, em Ciência da Computação as ontologias, que tiveram origem na comunidade de Inteligência Artificial (IA), são usadas para representação de conhecimento em computadores, visando a compreensão e o processamento automático de informação contextual. Na qualidade de bases de conhecimento, segundo [6] as ontologias servem para compartilhar conhecimento entre atores de um sistema, sejam humanos ou outros sistemas, e para aplicação de inferências e busca por informação, devendo integrar o conhecimento de um dado domínio, permitindo inclusive que o sistema possa aprender. Em relação a outras tecnologias, as ontologias ainda estão em fase relativamente recente de pesquisa e seu uso tem ganho cada vez mais popularidade. Eis algumas definições:

1. *"Uma ontologia identifica classes, cada uma caracterizada por propriedades que todos os elementos desta classe compartilham e as organiza hierarquicamente. Isto também inclui importantes relações entre classes e elementos, em um domínio de conhecimento específico",* [3].

2. *"Uma especificação explícita e formal de uma conceituação compartilhada",* [9].

Independentemente da representação usada, ontologias são formadas basicamente por dois componentes chamados *primitivas conceituais*, as quais são: *conceitos* e *relações*. Um conceito é uma classe onde se especificam atributos, qualidades ou propriedades comuns a todas as instâncias ou indivíduos desta classe. Por sua vez, as relações organizam e associam os conceitos. Existem vários tipos de relações, sendo mais corriqueiras aquelas que organizam os conceitos em hierarquia, especialmente hierarquia taxonômica, que é decorrente das relações de herança (relação classe/subclasse).

Quando as primitivas ontológicas são descritas apenas em *linguagem natural*, diz-se que a ontologia é *informal.* Ao adicionar Lógica na descrição das primitivas, a ontologia passa a ser dita *formal*, tornando-se essencial para a computação, por dar suporte à atuação de raciocinadores automáticos que se beneficiam de ferramentas lógicas para realizar inferências. Na abordagem de IA, os formalismos adotados para representar ontologias são *lógica de primeira ordem* [10], *frames* [13] e *lógica de descrição* [1].

1.2.1.1. Axiomas

Nas ontologias formais, as idéias de conceito e de relação vão ainda mais longe, pois estas primitivas são vistas como *axiomas*. Segundo a Lógica, axiomas são verdades auto-evidentes e que por isso não requerem prova, podem ser proposições assumidas conforme a conveniência ou podem ser regras e princípios universalmente aceitos. Os axiomas são o trunfo das ontologias em relação a sistemas de classificação comuns, pois os axiomas permitem a inserção de outras relações semânticas através do uso explícito de lógica. A idéia que temos sobre axiomas deve ficar bem clara e é a seguinte: *Os axiomas servem para caracterizar o ponto de vista que o autor da ontologia pretendia expressar ao criá-la.*

A representação dos axiomas influencia na medida e na qualidade em que a semântica axiomática poderá ser explorada por sistemas, bem como na portabilidade da ontologia, por isso é importante refletir sobre a *operacionalização* deles: a semântica provida por axiomas pode ser *formal* ou *operacional* [6]. A semântica formal é genérica e funciona como uma meta-semântica: assume que todos os axiomas são da forma "*antecedente⇒conseqüente", ou seja, se o antecedente é verdade, então o conseqüente também é verdade*. É uma semântica independente de domínio, que por isso garante a portabilidade da ontologia e da aplicação que a utiliza, mas que por outro lado pode restringir o entendimento das primitivas conceituais, já que é genérica e não necessariamente contempla e explora todas as nuances intrínsecas a um axioma. Por outro lado, na semântica operacional existe a necessidade de definir o *contexto de uso* de cada axioma dentro da aplicação, permitindo explorar intimamente o seu significado e desenvolver processos, métricas, ou seja, técnicas personalizadas para utilizar cada axioma, ao ponto destas técnicas serem totalmente inúteis a qualquer outro axioma e, muitas vezes, deixando os axiomas implícitos na execução sistema. Segundo [6], a *representação operacional de um axioma* é um conjunto de instruções, regras e/ou restrições para manipular o axioma e depende de um *cenário de uso* que descreve a maneira particular que os axiomas são usados para raciocinar. Por exemplo, especificamente para as relações de herança, existem diversas métricas desenvolvidas, muitas das quais definitivamente não se aplicam a outros axiomas. Como se vê, a semântica operacional favorece o domínio e integra seu conhecimento de maneira única, aumentando o entendimento das primitivas conceituais em detrimento da portabilidade da ontologia.

Por sua vez, os axiomas de uma ontologia dividem-se entre *axiomas de esquema* e *axiomas de domínio*. Os axiomas de esquema são genéricos e podem até funcionar como meta-axiomas, sendo bastante corriqueiros, já que, potencialmente, podem aparecer em todos os domínios de conhecimento. Conseqüentemente, é conveniente que muitos deles já venham integrados nas ferramentas para edição de ontologias. A utilidade de axiomas deste tipo é que a sua semântica é universalmente conhecida, o que, atrelado ao fato de serem corriqueiros, permite explorar o seu contexto de uso sem prejuízos à portabilidade da ontologia. São exemplos de axiomas de esquema as relações de herança, composição, exclusividade, incompatibilidade, equivalência e pertinência, além das abstrações, disjunções, cardinalidade e propriedades algébricas (reflexividade, simetria, transitividade, etc). Por outro lado, os axiomas de domínio podem não ser tão genéricos, pois visam delimitar apenas um ou poucos domínio e, se axiomas em geral ajudam o autor da ontologia

a expressar seu ponto de vista, os axiomas de domínio servem para expressar o domínio pretendido de forma ainda mais específica, íntima e direcionada.

Axiomas de domínio são raros, mas não pelo fato de serem pouco utilizados durante o desenvolvimento de uma ontologia, mas sim pelo fato de que eles são específicos e íntimos a um ou poucos domínios correlatos, e por isso dificilmente ele será encontrado em outros domínios diferentes, justificando a diferença entre os domínios. Logo, encontrar um axioma raro é uma grande evidência do domínio ao qual ele pertence e, quanto mais raro for, melhor será capaz de distinguir o seu domínio frente a outros domínios diferentes, bem como relacionar domínios correlatos. Exemplos de axiomas de domínio são: *O inimigo do meu amigo é meu inimigo*, *pai é um homem que tem filhos, todo deus é imortal*, *quem dorme cedo acorda cedo, a água se liquefaz acima de 0ºC,* etc. Como se vê, estes são axiomas dos mais diferentes domínios e seria muito difícil definir sempre o contexto de uso de cada um deles ao desenvolver uma nova ontologia. Além disso, não se pode esquecer que uma das grandes vantagens da engenharia ontológica é o reaproveitamento das ontologias nas mais diversas aplicações onde sejam relevantes, independentemente dos fins da aplicação. Daí a razão pela qual a portabilidade das ontologias é irrevogavelmente indispensável, levando à aplicação da semântica formal para representar os axiomas de domínio na grande maioria dos casos práticos, sem forçar a semântica operacional [6]. Além de tornar a ontologia independente da aplicação, isso normaliza os axiomas para que sejam submetidos a raciocinadores.

Usar axiomas é um investimento que trará ganhos em semântica: quanto mais axiomas adequados são usados numa ontologia, mais rica ela será e mais bem descrito será o seu domínio. Assim, quanto ao uso de axiomas, as ontologias são classificadas em *pesadas* e *leves*. As ontologias leves são pobres em axiomas e, praticamente, se restringem aos axiomas de esquema, principalmente as relações de herança, que são encontradas em qualquer ontologia. Já as ontologias pesadas são ricas tanto em axiomas de esquema quanto principalmente em axiomas de domínio. Por isso, as ontologias pesadas são mais úteis na medida em que trazem muito mais evidências para serem investigadas. Certamente, assim serão os recursos da Web Semântica: repletos de diferentes evidências. Sem dúvida, o estudo de evidências próprias a ontologias trará extensos benefícios a processos de mapeamento na Web.

1.2.2. MAPEAMENTO ENTRE ONTOLOGIAS

Dado que uma ontologia é uma abstração que representa conhecimento, identificando os conceitos e as relações de um domínio, então o mapeamento entre ontologias é o processo que identifica correspondências entre conceitos e relações de uma ontologia com os conceitos e relações de outras ontologias, caso estas correspondências existam.

Descobertas as correspondências, elas podem ser usadas para vários fins, desde a simples tarefa de serem exibidas até a tarefa de transformar uma ontologia em outra ou criar um conjunto de axiomas ponte entre as ontologias que funcionem como protocolo de comunicação, permitindo que as informações fluam entre as ontologias, caracterizando assim o compartilhamento de conhecimento.

Por exemplo, considere uma situação em que existem vários sistemas inteligentes isolados, cada qual com sua ontologia. Estes sistemas são capazes de interpretar o

conhecimento nas ontologias, no entanto são incapazes de se comunicar, criando ilhas de conhecimento. O ideal seria que existissem *pontes de comunicação* entre estas ilhas, permitindo o compartilhamento de conhecimento e, conseqüentemente, o fluxo de informação entre sistemas, ou seja, os sistemas seriam capazes de *interoperar* entre si. Logo, é fácil entender como que o problema da interoperabilidade entre sistemas recai na modelagem e no mapeamento entre ontologias e se aplica à questão de informações multiplamente representadas por várias ontologias, como no caso da Web.

Relacionar as primitivas conceituais de duas ontologias que compartilham o mesmo domínio de conhecimento simplesmente agrupando as primitivas mais similares não vale a pena se não for feito de tal maneira que preserve e respeite as estruturas lógicas entre as primitivas, bem como preservar e respeitar as interpretações pretendidas, como especificado pelos axiomas. Mapeamentos que assim fazem são chamados de *morfismos* [11], portanto daqui para frente toda referência feita a mapeamentos deverá ser entendida como um morfismo, pois a idéia deste trabalho é que todas as interpretações que satisfazem os axiomas da primeira ontologia também satisfaçam os axiomas da segunda ontologia após o mapeamento. Esse comentário é válido, pois teoricamente todos os mapeamentos deveriam ser morfismos, enquanto que, por exemplo, muitos trabalhos usam técnicas baseadas apenas em similaridade léxica e sintática e não garantem o morfismo, apesar de o terem como objetivo.

Segundo [4,5,7], dependendo das fontes de informação usadas nos mapeamentos, as *estratégias para investigação de evidências* de ontologias são classificadas como segue:

- *Investigação terminológica* – compara os rótulos das primitivas conceituais. Os rótulos são identificadores humanos, ou seja, são nomes e dependem do idioma: se *os rótulos são iguais, provavelmente as entidades também são*. As abordagens são (a) com base na dissimilaridade de termos, usando distância de edição, por exemplo, e (b) com base nas relações entre termos, considerando sinônimos como equivalências e hiponímias como subconceitos, por exemplo;

- *Investigação da estrutura interna* – compara evidências internas das entidades como, por exemplo, a cardinalidade dos atributos de conceitos e relações;

- *Investigação da estrutura externa* – investiga como os conceitos se relacionam com outros conceitos. Beneficia-se da representação de ontologias como grafos, podendo ser feita com base na árvore formada pelas relações de herança (taxonomia) ou com base no grafo formado por qualquer outra relação ou combinação de relações que possam inserir ciclos no grafo. Por exemplo, pode-se fazer comparação entre árvores ou *investigar a vizinhança*, que diz respeito a explorar a adjacência de um nodo, bem como a adjacência da adjacência e assim por diante, somando evidências internas e de extensões dos nodos vizinhos ou estipulando métricas de citação entre os nodos [2].

- *Investigação extensional* – compara as extensões conhecidas, entidades extras que não fazem parte mas se relacionam com a ontologia. O melhor exemplo neste caso são as instâncias dos conceitos;

- *Investigação semântica* – compara as interpretações das entidades, dado o contexto da ontologia. Aqui estão as abordagens que consideram axiomas e restrições.

De maneira geral, um algoritmo que implemente um mapeador automático recebe como entrada uma ontologia A, com n primitivas conceituais $\{a_0,a_1,...,a_n\}$, e uma ontologia B, com m primitivas conceituais $\{b_0,b_1,...,b_m\}$. O algoritmo deve produzir como saída a similaridade potencial entre cada a_i e b_j, indicada por casamentos representados como triplas da forma (a_i,b_j,Q_{ij}), onde Q é o qualificador do mapeamento. Veja:

Figura 1: Mapeador automático de conceito e relações ontológicas

Cada qualificador Q_{ij} é uma relação entre a_i e b_j, que depende da abordagem do mapeamento. Em mapeamentos sintáticos [8], Q é tomado simplesmente como um coeficiente de similaridade e $Q=\{x\in[0,1]\}$, onde a idéia é agrupar os pares a_i e b_j, com maior grau de similaridade. Note que quase todos os trabalhos existentes seguem abordagem sintática. Em mapeamentos semânticos [8], Q é tomado como uma relação semântica que funciona como axioma ponte. Por exemplo, em um sistema de hierarquia[1], as possíveis relações semânticas seriam $Q=\{\equiv,\supseteq,\subseteq,\neq,\cap\}$, sendo *equivalência* ($\equiv$) que significa $p\leftrightarrow q$, *mais geral* e *menos geral* ($\subseteq$, $\supseteq$) que significa analogamente $p\rightarrow q$, *diferença* ($\neq$) que significa $\neg(p\wedge q)$ e *sobreposição* ($\cap$) que significa uma interceção parcial.

A equivalência $(a_i,b_j,\equiv)$ é uma relação muito forte por dizer que a_i e b_j, são exatamente iguais. Mas a_i e b_j podem ser apenas parcialmente similares, talvez por um ser mais geral $(a_i,b_j,\subseteq)$ ou menos geral $(a_i,b_j,\supseteq)$ que o outro, relações as quais dão informação de herança, ou seja, pelo menos um conceito está contido no outro. A relação de sobreposição $(a_i,b_j,\cap)$ também indica similaridade, mas ela pode ser decorrente das relações $\subseteq$ e $\supseteq$ ou de uma simples interseção entre a_i e b_j, interseção da qual não se pode concluir nada de importante [8], pelo menos em princípio, já que nenhum dos dois conceitos está completamente contido no outro, dando margem para exceções.

[1] Para qualificar os mapeamentos, também poderiam ser consideradas outras relações semânticas além das de hierarquia.

Mesmo sendo o qualificador Q instanciado por relações semânticas, existe um coeficiente associado a Q [8], representando a probabilidade do mapeamento realmente estar qualificado pelo símbolo escolhido: a idéia é escolher a relação semântica que se mostrar mais forte, ou em outras palavras, ser mais confiável para cada par a_i e b_j,.

1.2.2.1. Classificação Automática e Ontologias Povoadas

A abordagem de muitos trabalhos é sobre o conceito de *ontologia povoada* [11], o que permite explorar a *relação de classificação* entre os conceitos e suas instâncias.

Dado que a *classificação das instâncias foi confiavelmente feita*, podemos esperar que esta relação preserve a corretude das estruturas lógicas das ontologias, facilitando intuitivamente a descobertas de relações de equivalência e de diferença: se dois conceitos compartilham as mesmas instâncias, é possível que sejam equivalentes; do contrário é possível que sejam diferentes.

Originado na comunidade de Aprendizado de Máquina e bastante popular na área de RI, o processo de classificação é um processo de *Aprendizado Supervisionado*, onde um sistema é treinado por um conjunto de dados confiavelmente classificados em uma base de dados dividida em classes conhecidas. Esta base de dados é chamada de *treino*, *treinamento, base de treino/treinamento ou conjunto de treino/treinamento,* constitui o conhecimento do classificador e não precisa estar formalmente estruturada. Baseado no treino, o sistema de classificação tenta identificar a classe correta de elementos desconhecidos, os quais constituem o *teste, conjunto de teste* ou *base de teste*.

Classificação automática não é novidade no ramo de mapeamento entre ontologias. Em [15], é esclarecida a utilidade de ontologias povoadas no mapeamento de ontologias, o que também é discutido em [11], onde são apresentados alguns trabalhos que fizeram uso desta abordagem. Em [12], afirma-se teoricamente que as instâncias dos conceitos ontológicos são uma boa evidência para relacionar conceitos que receberam interpretações dadas por comunidades diferentes.

Aplicar classificação automática para mapear ontologias da Web Semântica certamente será uma boa idéia, pois já existem muitos trabalhos que mostram sucesso ao classificar automaticamente recursos da Web atual. Em outras palavras, o estudo de evidências próprias aos recursos da Web tem trazido extensos benefícios a processos de classificação automática e certamente beneficiarão a abordagem ontológica da Web Semântica também, sugerindo adaptação das técnicas existentes.

Em [14] podem ser encontradas várias propostas de classificação de hipertexto que não utilizam apenas o texto do corpo das páginas da Web, mas também evidências diferentes, como o *título html* e o *conteúdo dos links* (*linkcontent*, em inglês). Em [2], são usadas abordagens sobre os *links* para extrair informação a partir da análise de evidências textuais seguindo a estrutura de *links* (*evidência de hipertexto*, também chamado *texto da vizinhança*), bem como análise sobre a maneira como as páginas se referenciam mutuamente através de seus *links* (*evidência de apontadores*) por meio da aplicação de elaboradas métricas para estimar a relevância das páginas em função da quantidade de referências. Estes trabalhos também servem para demonstrar como o uso combinado de diferentes evidências pode potencializar a precisão dos classificadores automáticos.

3. PROPOSTA DE TRABALHO

3.1. MAPEAMENTO SEMÂNTICO ENTRE ONTOLOGIAS UTILIZANDO AXIOMAS E CLASSICAÇÃO

Este trabalho propõe investigar formas de comparar as interpretações de ontologias e mapear conceitos, relações e axiomas de domínio, definidos pelo autor da ontologia. Ontologias que contemplam o mesmo domínio podem representá-lo de maneiras totalmente diferentes, contemplando até mesmo outros domínios ao mesmo tempo, por isso, em princípio, consideraremos ontologias sabidamente correlatas. O formalismo adotado para as ontologias será lógica de descrição (DL) e a linguagem para modelar as ontologias em DL será OWL (Ontology Web Language).

Assumiremos que uma fração significativa das ontologias é de ontologias pesadas. É interessante usar toda informação disponível na ontologia; uma vez que usamos apenas uma fonte de evidência, ficamos muito dependentes dela. Além disso, confiar apenas em evidências léxicas, sintáticas e de esquema pode ser perigoso pois não modelam a semântica do domínio. A combinação das fontes de informação, objetivando que o mapeamento final seja semântico, de fato pode garantir que muito mais entidades sejam corretamente mapeadas.

O uso de axiomas em mapeamento é novidade. Como o axioma é um dado que possui estreita relação com semântica, a pesquisa aqui proposta está centrada no uso de axiomas de domínio, bem como no uso da população da ontologia, beneficiando-se da relação de classificação e do sucesso das soluções em classificação automática. Isso não impede a utilização de rótulos, pois uma abordagem completa a outra: as evidências comuns a ontologias podem ser combinadas e assim pretendemos estender a informação de mapeamentos inicialmente terminológicos, segundo técnicas já existentes, para mapeamentos mais robustos com base nos axiomas e instâncias, investigando principalmente a idéia apresentada nos parágrafos seguintes. Portanto, este trabalho abrangerá três das cinco estratégias de comparação apresentadas na Seção 1.2.1: terminológica, extensional e semântica, com ênfase às duas últimas.

Antes de continuar, é interessante entender o que é um mapeamento semântico, ou seja, o que significa comparar interpretações. Para isso é importante entender as considerações a seguir. Analisando a partir do nível mais baixo, uma ontologia pode ser vista como uma abstração sobre um grafo $O=(V,E)$, onde V é o conjunto de vértices e E o conjunto de arestas. A cada vértice $v \in V$ existe um *rótulo* e um *conceito de vértice* associado. Por sua vez, o rótulo também possui um *conceito de rótulo* associado. Por exemplo, seja o vértice Melodias onde os conceitos de vértice e de rótulo são diferentes:

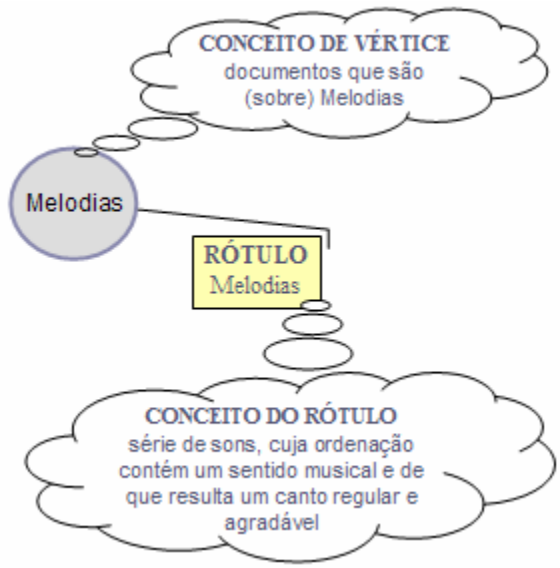

Figura 2: Conceito de vértice e conceito de rótulo

A observação trivial porém chave [8] é que rótulos em hierarquias de classificação são usados para definir o conjunto de documentos a serem classificados sob a classe que é identificada pelo rótulo. Logo, quando falamos de *Melodias*, não estamos querendo definir e nem nos referindo a "*uma série de sons, cuja ordenação contém um sentido musical e de que resulta um canto regular e agradável*", mas sim a "*documentos que são (sobre) Melodias*", esta sim é a semântica de classes onde as instâncias são documentos.

Como no exemplo, o conceito do rótulo não necessariamente é igual ao conceito do vértice, apesar de que estão relacionados [8]. O rótulo é apenas um identificador humano, uma facilidade, um termo que deve ser simples e pequeno para nomear o vértice, suficiente para desempenhar sua função que é terminológica e não de modelar lógica. Conseqüentemente, ao ser lido por uma pessoa, o rótulo remete à semântica mais comum que ele receberia no mundo real, semântica a qual constitui o conceito do rótulo. Porém, o conceito do rótulo de um vértice pode diferir significativamente do conceito do mesmo vértice, pois o conceito de vértice não é determinado pela terminologia mas sim pela lógica, dadas as relações com outros vértices.

É válido lembrar que as relações entre os vértices foram determinadas pelo autor, segundo o entendimento dele, em tempo de criação da ontologia. Isso significa que não devemos nos preocupar em criticar o que o autor quis dizer, como muitas pessoas tendem a fazer erroneamente, mas sim devemos entender o que o autor quis dizer: a função de um mapeador não é modificar e nem criticar uma ontologia, muito pelo contrário, ele tem de ser capaz de compreendê-la, de interpretá-la! Logo, guiar-se apenas pela semântica do rótulo pode nos enganar a primeira vista, prejudicando nosso entendimento de como fazer mapeamentos semânticos, na mesma medida como mapeamentos léxicos não são capazes de garantir mapeamentos semânticos.

Com base nas considerações acima, considere o exemplo seguinte (similar ao de [8]), onde dois grafos representam hierarquias de conceitos, num mapeamento entre sistemas de classificação hierárquica ou entre catálogos:

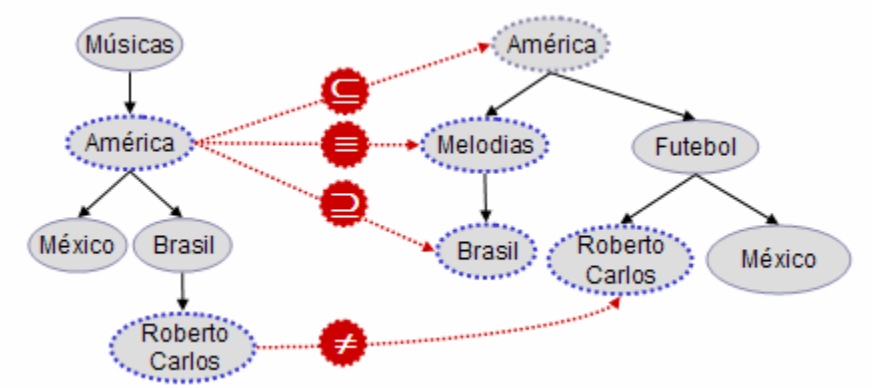

Figura 3: Exemplo de mapeamento semântico

Cada vértice tem um conceito associado, um rótulo associado e um conceito de rótulo associado. As setas pontilhadas representam relacionamentos *inter-ontológicos*, ou seja, representam mapeamentos e exemplificam quatro possíveis relações semânticas dentro de uma hierarquia. Por exemplo, o mapeamento (América, América, ⊆) é lido como *América é mais geral que América* e (Américas,Melodias, ≡) é lido como *América é equivalente a Melodias*. Os relacionamentos *intra-ontológicos* indicados por setas sólidas são do tipo *é um*, apesar do que se pensarmos a primeira vista na semântica dos rótulos, os relacionamentos não parecerão ser do tipo *é um*: *Roberto Carlos* ***não é um*** *Brasil*! Porém, os conceitos associados aos vértices *Roberto Carlos* e *Brasil* são, respectivamente, "*documentos que são (sobre) música de Roberto Carlos"* e "*documentos que são (sobre) Música do Brasil"*, sendo que "*documentos que são (sobre) Música de Roberto Carlos"* também são "*documentos que são (sobre) Música do Brasil",* que por sua vez também são "*documentos que são (sobre) Música"*. Veja que esta é a *interpretação pretendida pelo autor*.

Os vértices *América* e *Melodias* são associados como equivalentes no mapeamento (América,Melodias, ≡). Ora, mas porque foi feito este mapeamento se os rótulos são diferentes ao invés de (América,América,≡), por exemplo? Explicação: Apesar de terem rótulos diferentes, os vértices *América* e *Melodias* têm, subjetivamente, o mesmo conceito que é "*documentos que são (sobre) Música Americana"*. E mais, o mapeamento (América,América,≡) é inconsistente, pois o conceito do vértice *América* da ontologia direita é mais geral que o conceito do vértice *América* da ontologia esquerda, pois não se refere apenas a "*documentos que são (sobre) Música Americana"*, mas também a "*documentos que são (sobre) Futebol Americano"*, bem como poderia se referir a qualquer outro subdomínio de *América*, portanto a relação semântica mais apropriada e mais forte neste caso é (América,América, ⊆).

Por sua vez, o mapeamento (América,Brasil, ⊇) somado ao fato de que os vértices *Brasil* e *Brasil* têm o mesmo rótulo, é uma evidência de um possível mapeamento (Brasil, Brasil,≡)! Por fim, apesar de terem o mesmo rótulo, os vértices *Roberto Carlos* e *Roberto Carlos* possuem conceitos totalmente diferentes, logo a relação semântica mais forte é (Roberto Carlos, Roberto Carlos, ≠), veja: *Roberto Carlos* da esquerda tem o conceito "*documentos que são (sobre) Músicas e Roberto Carlos"*, enquanto que *Roberto Carlos* da direita tem o conceito *"documentos que são (sobre) Futebol e Roberto Carlos"*. Nada nesta

hierarquia de conceitos deixa claro se Roberto Carlos é uma pessoa e, se for, se trata-se da mesma pessoa ou não e, se não, se trata-se de um cantor e de um jogador: não se está tentando definir o que é um *Roberto Carlos*, mas sim generalizar o conteúdo de vários documentos que são sobre *Roberto Carlos*!

A idéia global a ser investigada neste trabalho utilizará a árvore de dependências entre os axiomas de domínio (não apenas a árvore de hierarquia), montada a partir da sua meta-semântica ou semântica formal "*antecedente⇒conseqüente*". Investigamos a hipótese de que quanto mais profundamente for definido um axioma α nesta árvore, ou seja, quanto mais específico ele for, podemos inferir duas coisas: (a) a comparação de α será mais fácil por depender de menos regras, pois mapeamentos inicialmente não semânticos do *conseqüente* tenderão a errar, desviar ou contrariar menos a verdade estabelecidas pelos axiomas do *antecedente* (b) por ser mais atômico, α tem grande potencialidade de compor outros axiomas, e por isso α terá maior possibilidade de ser encontrado em ambas ontologias sendo comparadas. Com isso esperamos selecionar axiomas prelinarmente mapeados que funcionem como evidências mais confiáveis para o mapeamento de outros axiomas, ou seja, dos axiomas dependentes.

Idealmente, os mapeamentos dos fatos devem ocorrer em primeiro lugar, pois eles são axiomas totalmente independentes de outros axiomas e a partir deles são definidos todos os outros axiomas. Portanto, o objetivo é traçar uma metodologia para estender mapeamentos de axiomas mais simples para mapeamentos de axiomas mais expressivos, respeitando as estruturas matematicamente lógicas da ontologia e, conseqüentemente, determinando mapeamentos semânticos livres de inconsistências, iniciando pelo mapeamento dos fatos ou por axiomas mais específicos quanto possíveis.

Veja outro exemplo, agora baseado não nas relações de hierarquia de um sistema classificado, mas sim na árvore de dependência entre os axiomas. O exemplo abrange o domínio das relações familiares e todos os mapeamentos destacados são equivalências:

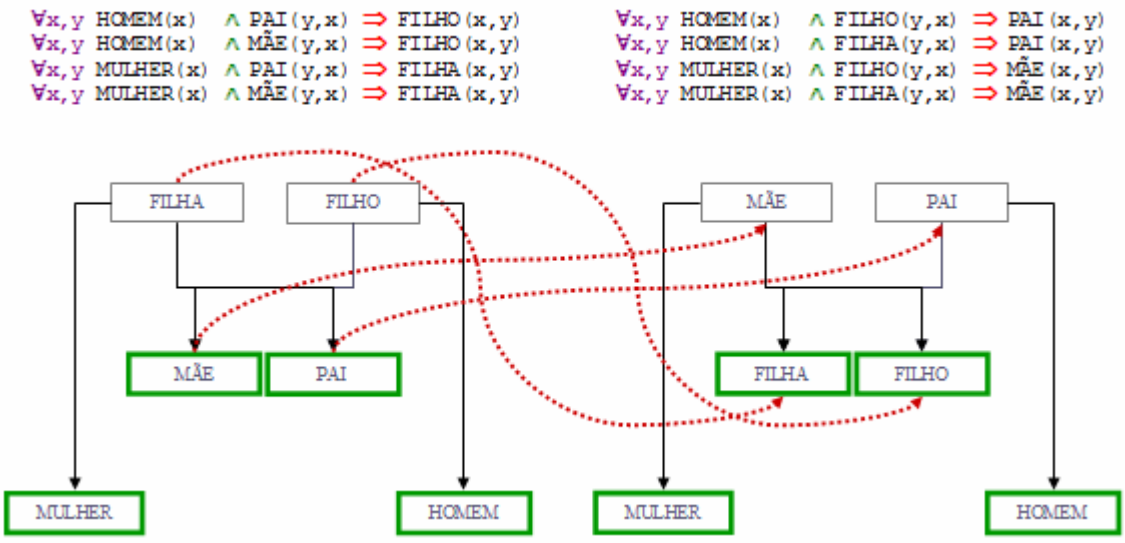

Figura 4: Exemplo de árvore de dependências entre axiomas

Este é um exemplo bastante simples de como conceitos podem ser definidos de diferentes maneiras e como conceitos mais específicos tendem a serem encontrados em ontologias que tratam do mesmo assunto. À exceção dos axiomas HOMEM e MULHER, os quais são conceitos, todas as relações são definidas de maneira diferente. Na árvore da direita, os axiomas FILHA e FILHO, os quais são relações, são definidos com base nos axiomas PAI e MÃE, que também são relações, mas são axiomas definidos como fatos. Os fatos PAI e MÃE também são encontrados na árvore da direita, mas não são definidos como fatos! Agora são definidos de maneira diferente em função de FILHA e FILHO, os quais agora também foram definidos de maneira diferente como fatos! Essa diferença é explicada pois as relações de paternidade e maternidade são inversas às relações de filiação. Apesar de tudo, os axiomas FILHO, FILHA, PAI e MÃE continuam sendo bastante específicos nas duas ontologias, mesmo porque eles são suficientes para definir várias outras relações familiares, como AVÔ, TIA, TIO, PRIMO, BISAVÓ, etc.

A abordagem terminológica não é a única maneira de obter mapeamentos iniciais. Como vimos, rótulos exatamente iguais ou muito parecidos podem se referir a conceitos ou interpretações totalmente diferentes e isso pode ocasionar muitos mapeamentos inconsistentes ao passo que é muito importante que os mapeamentos iniciais sejam confiáveis. Uma forma mais confiável de fazer isso é investigar as instâncias dos conceitos através de processos de classificação automática. Para isso precisaremos que as ontologias sejam povoadas (tenham instâncias) e que as suas instâncias sejam ricas em fontes de evidências, pois esta riqueza é um dos aspectos que aumenta a qualidade do classificador. Então, se a maioria das instâncias de um conceito de uma ontologia forem classificadas em apenas um determinado conceito de outra ontologia, provavelmente os dois conceitos são equivalentes, caso contrário são diferentes. Logo, o sucesso desta abordagem depende da qualidade do processo de classificação. Se usarmos os mapeamentos, obtidos através de classificação, para os conceitos mais simples, lembrando que conceitos são axiomas, podemos então estender estes mapeamentos para mapear axiomas mais expressivos seguindo a árvore de dependências entre os axiomas. Veja como, no exemplo a seguir, podemos descobrir (X,Y,≡) a partir dos mapeamentos (A,D,≡) e (B,C,≡):

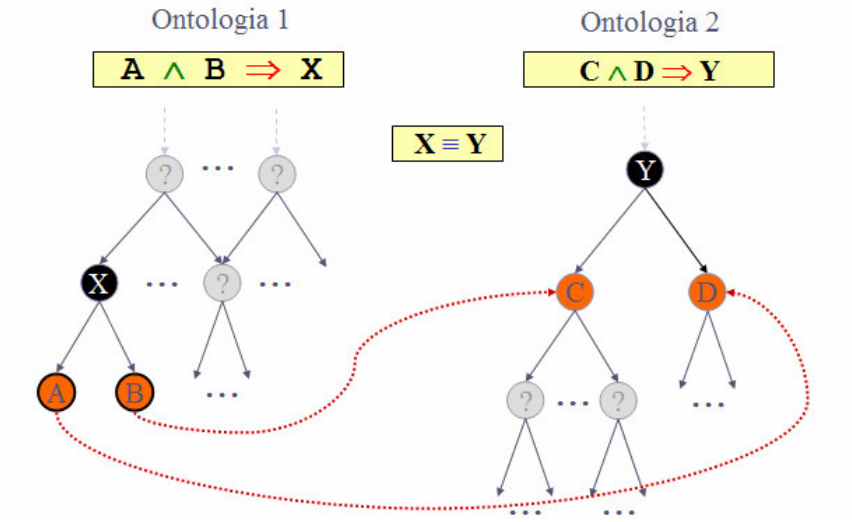

Figura 5: Exemplo simples de como estender mapeamentos

Recursos que representam instâncias classificadas e ricas em evidências são comuns na Web e sem dúvida serão comuns também na Web Semântica: já existe bastante pesquisa sobre classificadores automáticos de recursos da Web e já não é novidade que eles têm grande eficácia e eficiência ao lidar com hipertexto, muitos deles operando a cerca de 80% a 90% de precisão.

3.2. AVALIAÇÃO

O mapeador deve identificar automaticamente os pares a_i e b_j que de fato são correlatos, bem como deve indicar o qualificador Q correto que correlaciona o par. Mas a resposta do mapeador pode não coincidir com a realidade. Como discutido na secção 1.2.1, denotemos o casamento (a_i,b_j,QE_{ij}) como um *mapeamento estimado*, que é uma sugestão do mapeador para relacionar as entidades a_i e b_j,. Para avaliar a qualidade da decisão do mapeador, o casamento estimado é comparado ao casamento (a_i,b_j,QI_{ij}) que é o *mapeamento ideal* entre a_i e b_j, o qual deve ser previamente conhecido. Se o mapeamento ideal entre a_i e b_j não for indicado, pode-se inferir que ele seja $(a_i,b_j,\neq)$, ou seja, a_i e b_j não são análogos, não tendo qualquer tipo de similaridade. Espera-se que, dentre os mapeamentos estimados, exista a maior quantidade possível de mapeamentos ideais.

A avaliação de mapeadores é simples e bem parecida com a avaliação feita sobre qualquer sistema de RI. Considere que, para todo par de ontologias *A* e *B,* existe um conjunto de mapeamentos corretos e previamente conhecidos, que chamaremos de *conjunto ideal* ou *conjunto relevante,* e existe um conjunto de mapeamentos corretos ou incorretos produzidos como resultado de um mapeador, que chamaremos de *conjunto estimado* ou *conjunto computado* pelo mapeador: quanto maior a interseção entre estes dois conjuntos, melhor a qualidade do mapeador; logo, a avaliação consiste em comparar estes dois conjuntos. Dadas estas considerações, para quantificar a relevância de um conjunto de resultados, são usadas três métricas clássicas: *precisão, revocação* e *medida F*. Os valores destas métricas variam dentro da escala [0,1] e são definidas como segue:

$$precisão = \frac{|ideal \cap estimado|}{|estimado|} \qquad revocação = \frac{|ideal \cap estimado|}{|ideal|} \qquad medidaF = 2 \times \frac{precisão \times revocação}{precisão + revocação}$$

A precisão é uma métrica de corretude e por isso é afetada pela quantidade de respostas[2] irrelevantes, as quais são chamadas de *lixo*: quanto menos lixo nos resultados, maior a precisão do mapeador. O desejável é que a quantidade de respostas ideais supere a quantidade de respostas erradas. Logo, a precisão serve para medir a quantidade de acertos dentro do total de tentativas, ou, em outras palavras, a proporção de acertos em relação à proporção de erros.

Além de medir os acertos nas respostas, é desejável também que todas as respostas sabidamente ideais apareçam nos resultados: a abrangência dos resultados também é importante e é medida pela revocação que é uma métrica de completude; independentemente da quantidade de lixo que vier, deseja-se saber se é razoável a

[2] Entenda *resposta* como um mapeamento computado pelo mapeador sendo avaliado.

quantidade de respostas ideais dentro dos resultados. Logo, a revocação serve para medir se todos os acertos possíveis foram incluídos na resposta ou não.

Logo, para se obter valores máximos de precisão e revocação, o resultado deve conter todos os mapeamentos relevantes e apenas mapeamentos relevantes, quando então os conjuntos *ideal* e *estimado* serão exatamente iguais. Infelizmente, isso dificilmente acontece: é muito comum encontrar valores de precisão e revocação desequilibrados; ter uma precisão alta não é sinal de bons resultados se a revocação for baixa, e vice-versa. Para contornar o problema, costuma-se combinar precisão e revocação em apenas uma medida escalar, a medida F: quanto maior seu valor, mais equilibradas e maiores serão as medições de precisão e revocação.

Observe que, como estamos utilizando precisão, revocação e medida F para avaliar corretude e completude, significa que os métodos e técnicas desenvolvidos durante o trabalho serão verificados experimentalmente visando sua avaliação com respeito a sua *eficácia* (qualidade da resposta), pois ainda precisamos delinear um método. Portanto, o escopo de trabalho desta proposta ainda não visa otimização, ou seja, melhorar a *eficiência* (velocidade da resposta) do método. Mas nada impede que, caso comprovarmos sucesso para a eficácia, possamos considerar também um segundo trabalho para otimizar e avaliar a eficiência do método, o que se faz usando métricas que medem tempo. Por fim, serão feitas comparações de eficácia com outros sistemas e ferramentas apresentadas na literatura recente [8], no caso as ferramentas Cupid, COMA, SF e S-Match, dos quais o S-Match (Semântica Match) tem os mais altos valores de precisão, revocação e medida F.

3.3. CRONOGRAMA

1. *Revisão bibliográfica*: descobrir mais trabalhos relevantes na literatura.
2. *Aquisição de ontologias*: não possuímos uma base de ontologias de referência, portanto, para atingir nossos fins, ainda necessitamos adquirir ontologias prontas, pois não desenvolveremos as ontologias. As ontologias serão distintas porém pertencentes a domínios correlatos. Como queremos aplicar técnicas de classificação, temos como requisito que ontologias povoadas também sejam adquiridas.
3. *Técnica para os mapeamentos iniciais*: desenvolvimento de (a) técnica personalizada de classificação para mapeamentos iniciais de ontologias povoadas e/ou implementação de (b) técnica léxica conhecida (terminológica, por exemplo) para geração de mapeamentos iniciais de ontologias não povoadas. A escolha de qual(is) estratégia depende do sucesso em obter ontologias povoadas no ponto 2.
4. *Técnica para os mapeamentos com base em axiomas:* a partir dos mapeamentos iniciais, desenvolver e implementar mapeador de ontologias baseado nos axiomas (axiomas de domínio, a principio). Possivelmente este passo necessitará de um processo de padronização dos axiomas em um único formato (CNF, Cláusulas de *Horn*, etc...).
5. *Avaliação do mapeador*: avaliação do processo como um todo, mas também da estratégia de classificação desenvolvida.
6. *Redação*: redigir a dissertação e submeter artigo técnico a uma conferência científica reportando os resultados obtidos.

Meses	2007										2008		
	mar	Abr	mai	Jun	jul	ago	Set	out	nov	dez	jan	fev	mar
	1	2	3	4	5	6	7	8	9	10	11	12	13
Atividade													
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													

4. REFERÊNCIAS

[1] Baader, F.; Horrocks, I.; Sattler, U. "*Description Logics as Ontology Languages for the Semantic Web*". In Proceeding of the International Workshop on Ontologies, 2002.

[2] Calado, P.; Cristo, M; Moura, E. Et al. "*Combining Link-Based and Content-Based Methods for Web Document Classification*". In: X SPIRE/2003.

[3] Chandrasekaran, B.; Josephson, R.; Benjamins V. R. "*What Are Ontologies, and Why Do We Need Them?"*. IEEE Inteligent Systems, pages 20-26, January/February 1999, EUA.

[4] Euzenat, J.; Valtchev, P. "*An integrative proximity measure for ontology alignment*". In: Proceedings of Semantic Integration workshop at ISWC, 2003.

[5] Euzenat, J.; Valtchev, P. "*Similarity-based Ontology Alignment in OWL-Lite*". In Proceedings of the European Conference on Artificial Inteligence (ECAI'2004) , pages 333-337. IOS Press, 2004.

[6] Fürst, F.; Trichet, F. "*Axiom-based ontology matching*". In Proceedings of the 3rd International Conference on Knowledge Capture, Banff, Alberta, Canada, pages 195-196. October/2005.

[7] Fürst, F.; Trichet, F. "*Axiom-based ontology matching: a method and a experiment*". Relatório Técnico N° 05-02, Laboratório de Informática de Nantes-Atrantique (LINA), Março/2005. http://www.sciences.univnantes.fr/lina/fr/research/reports/

[8] Giunchiglia, F.; Shvaiko, P.; Yatskevich, M. *"S-Match: An Algorithm and an Implementation of Semantic Matching*". In Proceedings of the First European Semantic Web Symposium, pages 61-65. Springer-Verlag (LNC 3053), 2004.

[9] Gruber, T. R. *"Toward principles for the design of ontologies used for knowledge sharing"*. In: International Journal of Human-Computer Studies, v. 43, n. 5/6, p. 907-928, 1995.

[10] J. F. Sowa. *"Knowledge Representation - Logical, Philosophical and Computational Foundations"*. Brooks/Cole, Croton-on-Hudson, New York, 2000.

[11] Kalfoglou, Y.; Schorlemmer, M. "*Ontology mapping: the state of the art*". The Knowledge Engineering Review, v.18 n.1, p.1-31, January 2003.

[12] Kent, R. "*The information flow foundation for conceptual knowledge organization*". In Proceedings of the 6th International Conference of the International Society for Knowledge Organization (ISKO). Toronto, Canada, 10-13 July 2000.

[13] Minsky, M. "*A Framework for Representing Knowledge*". In P. Winston, editor, The Psychology of Computer Vision, pages 221–280. McGraw-Hill, New York, 1975.

[14] Yang, Y.; Slattery, S.; Ghani, R. "*A Study of approaches to hypertext categorization*". Journal of Inteligence Informations Systems, Special Issue on Automated Text Categorization, 2002.

[15] Y. Kalfoglou; M. Schorlemmer. "*Information-flow-based ontology mapping*". In On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE Lecture Notes in Computer Science 2519, Springer. Pages 1132–1151.