UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS
Instituto de Ciências Exatas
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
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 analogias entre ontologias para, em seguida, relacioná-las corretamente é uma tarefa de grande importância, visto que isto pode ser usado para que sistemas distintos se comuniquem em nível semântico, ou seja, interoperar entre si de maneira inteligente, similar ao que fazem seres humanos. Este trabalho propõe um estudo de como o uso de classificação automática e dos axiomas presentes nas ontologias pode evidenciar similaridade semântica, potencializando a precisão de mapeadores (os quais até hoje são fortemente baseados apenas em evidências léxica, sintáticas e na hierarquia de conceitos e relações), derivando métricas mais apropriadas tanto para descobrir mapeamentos quanto para qualificar cada mapeamento, associando-o com o operador lógico mais adequado, gerando assim axiomas de mapeamento.*

1. INTRODUÇÃO

1.1. MOTIVAÇÃO E OBJETIVOS

O mundo em que computadores sejam capazes de discernir e aprender é o sonho de muitos, dado que a automatização das 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 para completar 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 para imitar a cognição humana dentro de um computador, suprindo-o com competência suficiente para tomar decisões, situação na qual mesmo um humano pode se deparar com dúvida.

Descrever conhecimento através de ontologias vem se tornando cada vez mais popular, motivado 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 e migração das informações. Particularmente, as características ímpares da Web, como o seu imenso tamanho aliado ao crescimento exponencial do volume de dados, a volatilidade dos seus recursos, a diversidade de assuntos e a grande quantidade de usuários, levam a um desenvolvimento naturalmente distribuído de inúmeras ontologias para representar os recursos da 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), como no caso dos EIS (Enterprise Information System), sistemas de informação que auxiliam os executivos na tomada de decisão e que têm capacidade de *drill down* (refinamento), de *drill up* (síntese) e de filtragem de informações através de instruções de mapeamento definidas de forma totalmente manual. A este problema de integração ontológica chamamos de *casamento, alinhamento ou mapeamento entre ontologias.*

Em um certo ambiente, a formalização dos seus dados e informações visa diminuir sua heterogeneidade (desestruturação total ou parcial de conteúdo), 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, elas por si só inserem outro tipo de heterogeneidade em termos de semântica, dificultando os mapeamentos: uma mesma entidade pode receber nomes diferentes, bem como pode ser definida de maneiras distintas dentro de ontologias que descrevem o mesmo conhecimento ou, ao contrário, entidades com mesmo nome, mas que se referem a diferentes interpretações do mesmo nome dado o contexto de cada ontologia.

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 na prática é levar em consideração como axiomas são mapeados, mas a literatura é pobre em métricas que usam os axiomas; até a três anos atrás, não existiam trabalhos comparando as interpretações das entidades ontológicas, como bem dizem os trabalhos de [1] e [3], os quais concordam que o único algoritmo relevante até então era o trabalho de [5] e até hoje os trabalhos nesta área ainda são pioneiros. Dizer isso não significa que não existam trabalhos tratando de heurísticas para mapeamento entre ontologias, mas na prática, quase todos os trabalhos recaem na utilização de apenas algumas poucas evidências léxicas, sintáticas e taxonômicas para calcular um coeficiente de similaridade entre os elementos das ontologias envolvidas, associando como equivalentes os elementos que parecerem ser mais similares, sendo que esta forma de construir mapeamentos não garante respeitar e manter a estrutura lógica das ontologias envolvidas, ou seja, não garante a manutenção da semântica. Em suma, a explicação dada pelos autores é a seguinte [1]: *"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 do operador, será dada ênfase a classificadores e axiomas, 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 estruturas de dados usadas para representação de conhecimento humano em computadores, viabilizando a compreensão e o processamento automático de informação contextual, o que até hoje é feito apenas por humanos, pois são os únicos dotados de cognição e raciocínio. Na qualidade de bases de conhecimento, segundo [2] 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 todo 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 ganhado 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",* [7].

2. *"Uma especificação explícita e formal de uma conceitualização compartilhada",* [8].

Independentemente da representação usada, ontologias são formadas basicamente por duas componentes chamadas *primitivas conceituais*, as quais são: *conceitos* e *relações*. Um conceito é uma classe onde se especifica 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 Matemática e 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 todas as ferramentas matematicamente lógicas para criar inferências.

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, pois constituem o meio que capacita um computador a assimilar significado, ou seja, semântica.

A representação dos axiomas influencia na medida e na qualidade em que a sua semântica poderá ser extraída e então processada, 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* [2]. 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. Diferentemente, 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 rotinas e métricas personalizadas para cada axioma, a ponto de serem totalmente inúteis a qualquer outro axioma e, muitas vezes, deixando os axiomas implícitos na execução sistema. Segundo [2], 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 a partir da ontologia operacional. 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 aplicação. 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). Já os axiomas de domínio são bastante específicos: como o próprio nome já diz, eles servem para caracterizar intimamente o domínio que o autor da ontologia pretendia expressar ao criá-la. Não é de admirar que estes axiomas sejam naturalmente raros e quanto mais raros forem, melhor serão capazes de caracterizar e distinguir o domínio. 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 [2]. Além de tornar a ontologia independente da aplicação, isso normaliza os axiomas para que sejam submetidos a raciocinadores, como funciona com qualquer base de conhecimento.

Usar e abusar de axiomas é um investimento que trará ganhos em semântica: quanto mais axiomas 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 muito comuns 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: buscar por evidências do caráter semântico de recursos on-line é algo que deve ser levado em conta e bastante estudado

Portanto, após discutidos os argumentos acima, agora é possível entender que do uso de ontologias resultam indivíduos classificados em um sistema semanticamente organizado que inclui os relacionamentos naturais entre indivíduos.

1.2.2. MAPEAMENTO ENTRE ONTOLOGIAS

Dado que 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, pense numa situação em que existem vários sistemas inteligentes isolados, cada qual com sua ontologia. Estes sistemas são capazes de interpretar o conhecimento que têm sintetizado 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 e o aprendizado mútuo 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, ponto este discutido no capítulo 1.

Relacionar as primitivas conceituais de duas ontologias que compartilham o mesmo domínio de conhecimento, simplesmente agrupando os mais similares, não vale à pena se não for feito de tal maneira que preserve e respeite as estruturas matemáticas entre as primitivas, bem como preservar e respeitar as interpretações pretendidas, como especificado pelos axiomas. Mapeamentos que assim fazem são chamados de *morfismos* [6], 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. Segundo [1,3,4], as estratégias de comparação são classificadas como segue:

- *Comparação 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;
- *Comparação da estrutura interna* – compara, por exemplo, a cardinalidade dos atributos das primitivas conceituais;
- *Comparação da estrutura externa* – comparação com base em como as entidades se relacionam com outras entidades; investigação da vizinhança. 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.
- *Comparação Extensional* – compara as extensões conhecidas, entidades extras que não fazem parte mas se relacionam com a ontologia. O maior exemplo são as instâncias dos conceitos;
- *Comparação 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: Esquema de um mapeador automático de conceito e relações ontológicas

O qualificador Q do mapeamento é uma relação ou axioma ponte entre a_i e b_j, que depende da abordagem [5]. Em mapeamentos sintáticos, 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 anteriores são sintáticos. Em mapeamentos semânticos, Q é tomado como uma relação semântica representada por um símbolo, isto é, $Q = \{\equiv, \supseteq, \subseteq, \cap, \neq\}$, onde:

- Equivalência: ≡
- Mais geral e menos geral: ⊆, ⊇
- Diferença: ≠
- Sobreposição: ∩

A equivalência é a relação mais forte por dizer que a_i e b_j, são exatamente iguais; mais geral e menos geral dão informação de herança entre a_i e b_j,; a diferença indica relação entre a_i com o complemento de b_j ; a sobreposição não fornece informação muito importante.

Mesmo sendo o qualificador Q representado por símbolos, existe um coeficiente de confiança associado a Q, 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. Ontologias populadas

A abordagem de muitos trabalhos é sobre o conceito de *ontologia populada* [6], o que permite inserir a *relação de classificação* entre os conceitos ontológicos e suas instâncias: dado que a classificação das instâncias foi confiavelmente feita por especialista humano, podemos esperar que esta relação preserve a corretude das estruturas lógicas matemáticas das ontologias, facilitando intuitivamente a descobertas de relações de equivalência e de diferença: se dois conceitos compartilham as mesmas instâncias, é provável que sejam equivalentes; do contrário são 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 portanto chamada de *treino*, constitui o conhecimento do classificador e não precisa estar formalmente estruturada; baseado no treino o sistema de classificação tenta descobrir a classe correta de elemento desconhecidos que são chamados de elementos de *teste.* Classificação automática não é novidade no ramo de mapeamento entre ontologias: em [9], é esclarecida a utilidade e o uso de ontologias populadas no mapeamento de ontologias, o que também é discutido em [6], onde são apresentados alguns trabalhos que fizeram e fazem uso desta abordagem; em [10], é teoricamente discutido e afirmado que a população das ontologias é uma boa evidência para relacionar conceitos que receberam interpretações dadas por comunidades diferentes.

Particularmente na Web, o estudo de evidências próprias a seus recursos tem trazido extensos benefícios a processos de classificação automática e certamente beneficiarão a abordagem ontológica da Web Semântica também. Em [11] 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 [12] são usadas abordagens sobre os *links* para extrair informação a partir da análise de evidências textuais seguindo a estrutura de *links* (hipertexto, 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

Para o modelo deste trabalho, propõe-se abordar o problema dos mapeamentos semânticos. O grande objetivo é pesquisar e experimentar como comparar as interpretações de ontologias e mapear conceitos, relações e axiomas de domínio, definidos pelo autor da ontologia. Contudo, a maneira perfeita de fazer mapeamento semântico ainda é vaga. É difícil determinar exatamente quais entidades mapeiam entre si semanticamente, pois 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.

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 geral 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 exemplificar e entender o que é um mapeamento semântico, o que é comparar interpretação. Considere os dois grafos representando ontologias na figura abaixo:

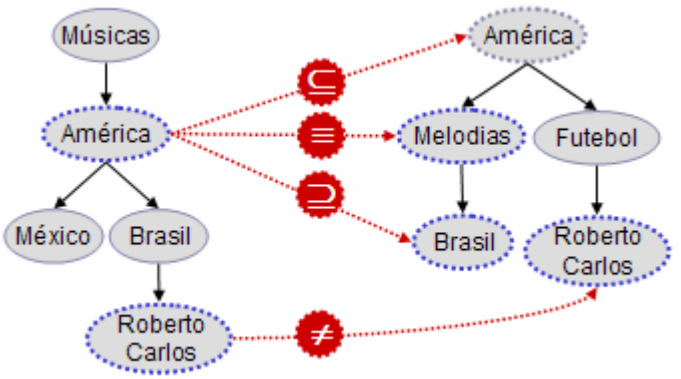


Figura 1: Exemplo de mapeamento semântico

As setas vermelhas pontilhadas representam relacionamentos *inter-ontológicos*, ou seja, representam mapeamentos e exemplificam quatro possíveis relações semânticas. Os relacionamentos *intra-ontológicos* indicados por setas pretas sólidas são do tipo *é um*. Observe que cada nodo do grafo tem um rótulo e um conceito associado. Se pensarmos apenas nos rótulos, os relacionamentos não parecerão ser do tipo *é um*: "Roberto Carlos" ***não é um*** "Brasil"! Porém, o conceito associado aos nodos "Roberto Carlos" e "Brasil" são, respectivamente, "*música de Roberto Carlos*" e "*música do Brasil*", sendo que "música de Roberto Carlos" ***é uma*** "música do Brasil" . Veja que esta é a *interpretação pretendida pelo autor*.

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

Por sua vez, o mapeamento (América,Brasil, $\supseteq$) somado ao fato de que os nodos "Brasil" e "Brasil" tem o mesmo rótulo, é uma evidência de um possível mapeamento (Brasil, Brasil,$\equiv$)! Por fim, apesar de terem o mesmo rótulo, os nodos "Roberto Carlos" e "Roberto Carlos" possuem conceitos totalmente diferentes: "Roberto Carlos" da esquerda se refere às músicas brasileiras cantadas pelo artista Roberto Carlos, enquanto que "Roberto Carlos" da direita se refere ao futebol praticado pelo atleta Roberto Carlos, que é outra pessoa, e, mesmo que fossem a mesma pessoa, um cantor que também é jogador de futebol, o objetivo dos dois nodos continua sendo diferente, portanto seu contexto, seu conceito também é diferente. Logo, estes nodos não podem ter mapeamento qualificado nem como $\supseteq$ ou $\subseteq$: a relação semântica mais forte é (Roberto Carlos, Roberto Carlos, $\neq$).

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*$\Rightarrow$*conseqüente"*. Partimos do pressuposto de que quanto mais profundamente for definido um axioma α nesta árvore, ou seja, quanto mais específico ele for, podemos inferir duas coisas: acreditamos que (a) maior será a facilidade de comparar α, por depender de menos axiomas e por isso ter menos compromisso/impacto nas estruturas matemáticas, tornando a comparação da sua interpretação menos sujeita a erros lógicos e (b) por ser mais atômico, α tem grande potencialidade de compor outros axiomas, e por isso α terá maior probabilidade de ser encontrado em ambas as ontologias 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. Em outras palavras, se pudermos garantir a confiabilidade do mapeamento do *antecedente*, também poderemos garantir a confiabilidade do mapeamento do *conseqüente.*

Se radicalizarmos esta idéia, veremos que o ideal é ter os mapeamentos dos fatos 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 tanto quanto possível.

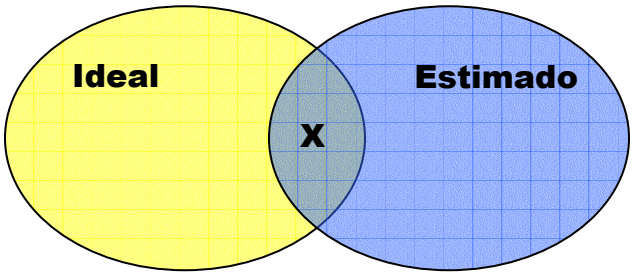
A terminologia não é a única maneira de obter mapeamentos iniciais. Como vimos, rótulos exatamente iguais ou muito parecidos podem se referir a conceitos ou interpretações totalmente diferentes. Uma forma mais confiável de fazer a mesma coisa é investigar as instâncias através de processos de classificação automática. Para isso precisaremos que as ontologias sejam populadas e que as suas instâncias sejam ricas em fontes de evidências, o que é justamente o caso da Web e sem dúvida será da Web Semântica também: já existe bastante pesquisa sobre classificadores automáticos de recursos da Web e já não é novidade que eles tem grande eficácia e eficiência ao lidar com hipertexto, muitos deles funcionando a cerca de 80% a 90% de precisão.

Os métodos e técnicas desenvolvidos durante o trabalho serão verificados experimentalmente visando sua avaliação empírica com respeito a sua eficácia. Se comprovarmos a eficácia como positiva, consideraremos também otimizar e avaliar também a eficiência do método.

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. 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 $(a_i,b_j,\perp)$, 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 classificadores automáticos, bem como sobre qualquer outro sistema de RI. Para todo par de ontologias A e B existe um conjunto de mapeamentos corretos que chamaremos de *conjunto ideal* e um conjunto de mapeamentos corretos ou incorretos produzidos pelo mapeador que chamaremos de *conjunto estimado*. Quanto maior a interseção X entre estes dois conjuntos, melhor a qualidade do mapeador. Veja:



Para quantificar relevância de um conjunto de resultados produzidos por experimentos de RI, são usadas três métricas clássicas: *precisão, revocação* e *medida F*. Os valores destas métricas variam dentro da escala [0,1].

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

A precisão é uma métrica de corretude, é afetada pela quantidade de lixo ou respostas irrelevantes: quanto menos lixo nos resultados, maior a precisão. 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 seja, 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 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 objetos relevantes e apenas objetos relevantes, quando então os conjuntos *ideal* e *estimado* serão exatamente iguais. Infelizmente, isso dificilmente acontece e 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 é sinal de que mais equilibradas e maiores serão as medições de precisão e revocação.

Por fim, serão feitas comparações com outros sistemas e ferramentas apresentadas na literatura recente [5], no caso as ferramentas Cupid, COMA, SF e S-Match.

3.3. CRONOGRAMA

1. Revisão bibliográfica.
2. Aquisição de ontologias bem modeladas e distintas, porém análogas, ou seja, pertencentes ao mesmo domínio; buscar por ontologias satisfatoriamente populadas.
3. Desenvolvimento de técnica de classificação personalizada para geração de mapeamentos iniciais de ontologias populadas e/ou implementação de técnica terminológica conhecida para geração de mapeamentos iniciais de ontologias não populadas. A escolha de qual implementação depende do sucesso em obter ontologias populadas no ponto 2.
4. Desenvolvimento de um mapeador de ontologias baseado nos axiomas, principalmente nos axiomas de domínio e hierarquia de primitivas conceituais, que partirá do mapeador inicial, desenvolvido no passo anterior. Possivelmente necessitará de um processo de padronização dos axiomas em um único formato (CNF, Cláusulas de *Horn*, etc...).
5. Avaliação do mapeador, inclusive do processo inicial baseado em classificação de ontologias populadas.
6. Redigir a dissertação e submissão de 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] Fürst, F. e 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/

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

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

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

[5] J. Euzenat and P. Valtchev. *"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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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).

[11] 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.

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