Integração de dados baseada em ontologias

Data integration based on ontologies

Maurício Barcellos Almeida[a]

Transcrição da apresentação

Eu vou falar um pouco sobre a interação de dados baseados em ontologias que é um interesse de pesquisa já há uma década. Como foi falado eu sou da UFMG, do Programa de Pós-Graduação em Gestão e Organização do Conhecimento, da Ciência da Informação, o grupo de pesquisa também liderado por mim e pelo professor Renato da FGV, que é o "Representação do Conhecimento e Ontologias de Dados", e tenho trabalhado também em conjuntos com universidades estrangeiras com pesquisas nessa área de modelagem e ontologias.

A palestra está dividida em quatro partes, a primeira é mais uma introdução, e o objetivo é explicar um pouquinho o que é interoperabilidade e como a ontologia pode ajudar nessa questão especialmente porque ontologia pode resolver tanto problemas de interoperabilidade quanto problema de ter solução. Mas se a gente souber identificar tipos de dados que a gente está manipulando talvez seja mais fácil aplicar as ferramentas adequadas inclusive a ontologia, é isso que eu pretendo explicar para vocês.

O que é ontologia? é uma palavra que muita gente tem dúvida, se você colocar no Google vai achar desde Filosofia até Computação, mas é uma palavra bem mais conhecida do que a gente pensa. Aqui por exemplo tem a porta de um D.A. lá da UFMG escrito "uma casa ontologicamente sustentável", não tenho certeza se a pessoa que escreveu sabe do que se trata mas está lá.

Aqui também tem outro exemplo de uma pichação na época das manifestações de 2013 e 2014, isso é lá no Rio Grande do Sul, um aluno de lá tirou e enviou para a gente. Mas o que realmente define melhor ontologia, por incrível que pareça, é o Neymar. Aqui nós temos o Neymar porque um dom ontológico, logicamente isso é uma brincadeira de um jornal de Belo Horizonte que custa 25 centavos, o que a gente pode esperar que seja de uma informação fidedigna. Mas é interessante ver com é que a ontologia está em todos os cantos.

Saindo da brincadeira, a ontologia é um termo que aparece desde a Filosofia Alemã desde o século 17, quase há dois séculos, e hoje em dia a gente encontra também muito área da Computação e da Ciência da Informação, e por um órgão que regula a tecnologia na área de Web Semântica.

A melhor definição de ontologia que eu consegui até hoje, apesar de várias que existem, é uma definição até antiga de um autor na área de inteligência artificial que fala que é tipo um cabide conceitual, onde nós vamos entulhando conceitos que podem ser usados para diversos motivos, mas a estrutura principal é essa esquelética, é o que a gente pode depois preencher com mais dados.

a Universidade Federal de Minas Gerais (UFMG). E-mail: mba@eb.ufmg.br. Currículo: http://lattes.cnpq.br/5218069708058487

A ontologia aplicada de certa forma ela vai reunir todos esses conceitos da metafísica e ontologia filosófica e aplicar como princípio de modelagem, a construção de ontologias na área da Computação e da Ciência da Informação. Então basicamente ela abre com conceitos e a gente vai descrevendo o mundo.

Nesse aspecto, da mesma forma que as pessoas amam Saracevic, a palavra lá em 96 na ligação com a inteligência artificial, do mesmo jeito que as pessoas precisam dos conceitos para falar e se expressar, o computador precisa dessa estrutura de mundo para fazer inferências automáticas.

Os aspectos que a gente pode também auxiliar hoje em dia é a questão de interoperabilidade dos sistemas, principalmente os sistemas médicos. A interoperabilidade é uma palavra que está na moda, todo mundo fala da interoperabilidade, mas realmente é uma coisa que a gente escuta mais falar do que uma solução.

Na verdade, eu acredito que não há uma solução perfeita, acredito que a gente pode mitigar o problema. No caso dos sistemas médicos, como a gente tem trabalhado lá na UFMG, por exemplo, na área de obstetrícia, tem dificuldades porque tem consultas médicas desde o parto até depois a continuidade dos cuidados ao recém-nascido, e isso é importante que o profissional de medicina possa ter acesso aos prontuários dessa pessoa independente de onde ela está, que podem ser hospitais diferentes mas tudo tem que estar transparente. Então a ideia da Medicina está muito ligado a essa questão com a nossa pesquisa que pretende ajudar.

Falando um pouquinho sobre a representação do conhecimento então, do ponto de vista semiótico, eu vou até começar a falar mas deve ter especialistas aí na área de semiótica, mas eu vou ser bem simples até porque tem pouco tempo. Mas a semiótica é a forma que o significado é gerado através da interpretação de dados sensoriais, e o criador foi Pierce, matemático e filósofo do século 19.

Peirce criou um triângulo conhecido e que tem uma lógica criada pelo próprio Peirce, e depois disso a gente teve diversas interpretações na área de Ciência da Informação, e também na década de 70 mais conhecida na nossa área de Biblioteconomia e Ciência da Informação.

Qualquer que seja o triângulo adotado a ideia da semiótica é bem simples, a pessoa a medida que ela percebe pelo signo, desenho ou palavra escrita, ela forma um conceito na cabeça dela a partir da lembrança que ela lembra de ter visto do objeto. Então esse desenho do elefante ela lembra que o pai levou no zoológico quando era pequena e lembra que esse bicho aqui é um elefante.

A ideia do triângulo semiótico, agora adaptado para a ontologia, eu tenho aqui um objeto que pode ser representado por um signo a partir de uma compreensão, então eu chamo aquele animal de elefante porque eu aprendi assim assim, é uma convenção. No entanto, eu poderia chamar gato se fosse assim que fosse ensinado a gente desde pequeno.

O conceito que se forma em nossa cabeça a partir da experiência, aquela do signo, e também temos as arestas do lado direito para a percepção, quando eu vejo um signo eu formo o conceito. Segundo o Pierce é isso que acontece com um signo e como interpretar o que ocorre no ambiente.

Eu gostaria de pedir licença para sugerir um novo triângulo um pouquinho diferente, que vai ficar mais fácil de identificar o que eu quero passar para vocês. Então esse triângulo de significados que eu vou apresentar aqui. Desse lado eu vou ter Símbolos, que pode ser diverso, pode ser cachorro, dog, perro, outros nomes de cachorro, são todas as coisas que vão se referir aquele animal cachorro. Desse lado eu tenho os indivíduos, que eu coloquei uma foto do Marley que é um cachorro específico, e desse lado eu vou pedir licença para colocar um outro conceito que é o Universal. O Universal que é muito controverso no plano aristotélico, basicamente ele quer dizer um título natural, que existe independente da nossa vontade, por exemplo, aquela árvore ali fora existe, o sol existe, se por acaso eu morrer amanhã o sol continua lá.

A classe é um pouquinho diferente. A classe é uma criação humana para determinado fim, por exemplo, eu crio a classe das pessoas que estão assistindo aula tal hora porque tem que fazer chamada, a classe dos carros que param no estacionamento da UFPB porque vão ter um crachá, então é um pouquinho diferente dos Universais mas até as pessoas tratam como se fosse a mesma coisa.

Nesse triângulo eu tenho também os símbolos, representando indivíduos, a aresta da denotação, os símbolos também remetem aos universais, e os indivíduos instanciam do universal. Eu tenho um universal cachorro eu tenho Marley que instância o universal cachorro, mas eu posso criar a classe cachorro que vão ser vacinados.

Com esse triângulo eu vou dar alguns exemplos que vão me ajudar explicar minha ideia. Então eu tenho aqui o indivíduo o Marley, que é o cachorro, outro exemplo, Marley é um cão, no lado aqui dos símbolos eu tenho que cão é um substantivo, já não estou falando de indivíduo, estou falando de entidade linguística, palavra cão. Cannis familiaris então são sinônimos, um outro aspecto linguístico, a questão da sinonímia. Cães são vertebrados é uma propriedade universal, todos os cães são vertebrados, não existe cão que não seja vertebrado. Por último eu tenho que cães são possíveis transmissores de raiva, então essa é uma possibilidade, é um conhecimento contingente.

Com esse exemplo eu vou criar o que eu vou chamar de níveis de representação, e daí vou dar vários exemplos de como a tecnologia pode atuar nesses níveis de forma a contribuir com o controle a atividade.

Os níveis que eu vou apresentar para vocês eu vou chamar de nível ontológico, de contingente, linguístico e pactual. Aqui eu apresento de novo a figura que acabamos de ver, aqui do lado dos indivíduos nós temos o nível pactual, com declarações das entidades e seus relacionamentos com o mundo, que é o caso do Marley que vive em um lugar chamado Terra, o Marley é o cão.

Do lado de cá eu tenho as propriedades dos significados para os signos de linguagem, é uma especialização do signo, aqui eu estou me referindo às questões linguísticas, são os substantivos, eu tenho sinônimos. Em cima eu estou chamando de nível ontológico, são signos universalmente verdadeiros, tem vários exemplos disso na Medicina que nós estamos lidando, só temos hepatite no fígado, não tem como ter hepatite em outro lugar. São vários axiomas que existem e podem ser facilmente levantados. Por último o nível do contingente, aqui o exemplo que cães são possíveis transmissor de raiva.

Eu vou mostrar agora para cada um desses níveis exemplos de como a tecnologia pode agir e ao final a gente vai mostrar como os sistemas vão usar esses diferentes níveis, então se a pessoa entender o nível de dados que ela está tratando ela vai poder utilizar a ferramenta tecnológica mais adequada, a gente vai ver que a possibilidade de conseguir seria via ontologias.

Um exemplo primeiro do nível pactual, são declarações sobre indivíduos, no caso aqui sobre o cachorro Marley. Declarações sobre indivíduo são feitas na computação desde os anos 70 pelos especialistas pela tripla entidade-atributo-valor, ou E-A-V. Nos anos 2000 a gente teve coisa parecida que foi inventada no caso o RDF falado na palestra anterior, e que não é nada diferente do que foi criado nos anos 70, o que muda agora é que se chama sujeito – predicado – objeto, mas é a mesma ideia.

RDF já tem quase 20 anos, tem outros recursos bem mais avançados para lidar com dados, mas RDF continua sendo algo muito importante. Então essa linguagem sujeito – predicado – objeto é exatamente isso, Marley é um tipo de cachorro, Marley viveu na Terra, então não tem uma forma mais intuitiva de apresentar dados do que o RDF. Me parece bem mais simples que os bancos de dados, mas, de qualquer forma, fica a ressalva que ela tem limitações.

A gente precisa extrair dados sobre os indivíduos, no caso la o Marley, mas aqui eu tenho um exemplo de como se fosse um prontuário médico de uma pessoa com alguma doença e eu tenho que instanciar essa pessoa com um tipo de doença e para um termo geral que está em um vocabulário. Então essa extração dos termos do prontuário são marcados com esses termos podem ser identificados por outros sistemas. Essa é uma forma de extração da informação no ponto de manipulação de indivíduos.

Agora vou dar um exemplo sobre a questão dos signos, do nível do bicho. Aqui uma marcação das partes do discurso que mostra que aquele erro lá é um substantivo, o fato dele ser um substantivo vai ajudar muito aos programas aí que lidam com padrões léxicos e lidam com uma série de coisas então essa marcação é muito importante, e o processamento da linguagem natural fazem isso com muita eficácia hoje em dia.

Aqui um exemplo também de um padrão para extrair informações de um determinado focus, por exemplo se usar dados de padrões eu consigo definir que CDA significa alguma coisa, então a gente vai pegar os prontuários existentes dessas formas. Uma coisa que eu esqueci de comentar ali é que por mais que a gente goste de estruturas em um prontuário de pacientes, os médicos quase sempre prefere escrever como um texto, é o que a gente tem que tentar para ajudar o profissional então essa questão do texto livre apesar de ser inicialmente uma coisa muito utilizada, seria forma preferível dos profissionais que vão lidar com isso.

Um outro exemplo no nível ontológico, dos universais. Ali a gente tem uma representação em OWL, uma linguagem bem mais avançada que o RDF, e eu tenho aqui cachorro que é subclasse vertebrada, vertebrado que é subclasse de animal, e a classe dos vertebrados quando eu coloco palavras equivalentes a eu to colocando necessários o suficiente, que é o que se deve fazer com o conceito completamente, então eu estou colocando vertebrado em uma classe que o animal tem alguma parte. Essa é uma lógica meio aportuguesada mas se eu colocar isso em OWL em um motor de inferência como o Protege ou de ontologias, eu vou achar que é um assunto universal no questionário, e a ontologia é capaz desse tipo de coisa.

A ontologia não vai ser uma panaceia que vai resolver todos os problemas de integração, ela vai resolver problemas com nível básico, problemas em um nível mais abstrato, mais próximo à necessidade humana vão ser inicialmente difícil de resolver, mas é importante que a gente comece com algum nível de dificuldade e depois vão evoluindo.

Aqui um exemplo dessa inferência que eu acabei de mostrar, não existem pontos que não tenha ossos, então eu tenho ali a classe vertebrada embaixo de animal. Protege pra quem não conhece é um construtor de ontologia que vem com um motor de inferência, que são algoritmos que fazer inferências automáticas.

São coisas simples mas são exemplos de inteligência artificial, então eu tenho aqui vertebrado ao lado direito a classe equivalente tem em marte alguma vértebra, subclasse animal. Se eu rodar uma consulta como esta no lado direito ali e peço todos os cachorros que não tem alguma parte que não seja osso eu vou ter inferência e vou ver ali no lado direito que não existe essa classe. Então você construir ontologias na Ciência da Informação e trabalhando com vocabulário, é muito interessante porque o motor de inferência ajuda a achar coisas que você não vê. Às vezes você vai trabalhar com um vocabulário é bem razoável você conferir ali a consistência do seu vocabulário.

Aqui tem outro exemplo de extração de relações taxonômicas, que são oscilações básicas que a gente encontra em ontologias, através de chavões taxonômicos. Então eu tenho se ele acha um substantivo, como aqui no exemplo, então ele vai identificar que esse GAR 3 vai é subclasse de uma proteína. Abaixo outros padrões que eu vou conseguir identificar que fraturas são tipos de machucados, e por exemplo, um tipo de doença mental.

Aqui um exemplo de extração de outras relações, por exemplo, o algoritmo que faz o web mining procurando na própria Internet ou banco de dados Wikipédia ou outros DBPedia, e, por exemplo, aqui vamos ter que achar um tipo de informação buscando, então estudar sua definição que está no seu lado direito que foi buscar na própria Web e conseguiu a informação.

Por último, o exemplo do nível do contingente, que seria o conhecimento possível, tem um exemplo desses vetores de raiva. Essa aqui é uma fala de um importante cientista de ontologia do mundo, é um professor de Manchester, ele fala que muito trabalho sobre ontologia na área da computação tem objetivo integrar o raciocínio tipo com o raciocínio probabilístico, isso porque Medicina trabalha grande parte com dados estatísticos, mas a ontologia trabalha apenas com raciocínio clássico, então em geral, sistemas clínicos vão muito além do que ontologias podem proporcionar. Mas a ontologia é o primeiro nível de integração para sistemas para depois pensar em níveis mais avançados e abstratos próximos do raciocínio humano.

Um exemplo da tripla RDF para mostrar como é que conclusões podem ser representados em RDF. Eu já tinha dito que o RDF não é uma linguagem com semântica bem definida, os termos podem ter diferentes interpretações, eles são complexos no sentido de que são montados, e não posso afirmar 100% dos casos quando faço declarações RDF.

No exemplo que eu dei sobre a tripla sujeito – predicado – objeto, eu tenho, por exemplo, na segunda linha que tabaco causa câncer, a gente sabe que isso é verdade mas não é verdade em 100% dos casos. Então um caso representação que pode ser feita em RDF e outros casos mas que se tem que o cachorro pode ser transmissor de raiva mas nem sempre.

Um vocabulário muito conhecido da Ciência da Informação é o Mesh, que foi criado por bibliotecários e traz também exemplo dessa ideia de conhecimento contingente com só ocorrências, por exemplo, e se o conceito que está definido aqui sobre o bipolar ele tá definindo no Mesh como uma terapia de drogas, como complicação, então tem vários labels que eu posso classificar, demonstrando várias possibilidades para a ação de dados.

Por final as diretrizes que são árvores de decisão, para casos de urgência ou bem conhecidos, que um médico vai seguir determinados padrões, inclusive de aplicações éticas, se ele não segue esses padrões depois têm que se explicar porque não seguiu as diretrizes do hospital. Então a gente pode tirar várias tags que são utilizadas nos sistemas de construção de ontologias, baseados no conhecimento.

O eu passei para vocês uma grande quantidade de informações em um período relativamente pequeno, aquela teoria geral que a gente viu desde o início que tinham ali os quatro níveis, vieram lá do triângulo do Peirce originariamente, a gente está fazendo essa conexão, e representamos esses quatro tipos de níveis de dados, o linguístico, o nível pactual, o contingente e o ontológico.

Quando a gente junta tudo isso agora é possível ver que a gente tem de um lado, o lado esquerdo aqui os signos, conhecimento sustentável, quer dizer, uma vez que eu definir que cão é sinônimo de canis, aquilo ali está resolvido, e do lado direito a gente tem o conhecimento dinâmico que vai representar as coisas que acontecem no mundo na medida que vão conhecendo.

Então do lado do conhecimento sustentável está o conhecimento ontológico, são axiomas fixos, mas obviamente é falível, a ciência está sempre evoluindo então como um livro tem que ser atualizado em 10 anos a ontologia também. É mais para mostrar como a gente tem que pensar a ontologia como uma ferramenta que carece de ser sempre atualizada e a gente chama de curadoria.

Agora eu vou então ligar todos esses níveis e mostrar a minha ideia de como é que os sistemas de informação deveria funcionar de maneira geral, e concluir como as ontologias podem ajudar nesse sentido.

Aqui nós temos os usuários de um sistema de informação e aqui eu coloquei dados em repositórios que representam aqueles níveis de conhecimento que eu acabei de mostrar, como repositório de signos vai ser de ferramentas pactual, as ontologias que são modelos da realidade, repositório de regras, por exemplo as diretrizes que eu mostrei, e repositório de fatos consistentes com a realidade, como é o caso do pão né, então eu tenho todos esses níveis dispersos e misturados das entidades que vou preparar.

Os sistemas de informação vão utilizar todos esses níveis, vão usar o motor de regras de fato, as ontologias e os signos. Os signos vão representar as ontologias que propriamente mostram o que é a realidade e também vão ser utilizados no sistema de informação onde a gente ter a interface de acesso ao usuário. Já as ontologias vão conter o significado das regras para os repositórios de fatos, e você articulam as coisas que acontecem na prática. Dessa forma, com essa arquitetura, eu vejo a forma que as ontologias podem contribuir na realidade.

Primeiro eu queria dizer que a interoperabilidade semântica, com a conotação de semântica formal da Web Semântica, não se trata da semântica que a gente trabalha na linguagem natural, e as ontologias podem ajudar a mitigar esse problema da interoperabilidade em um nível realmente

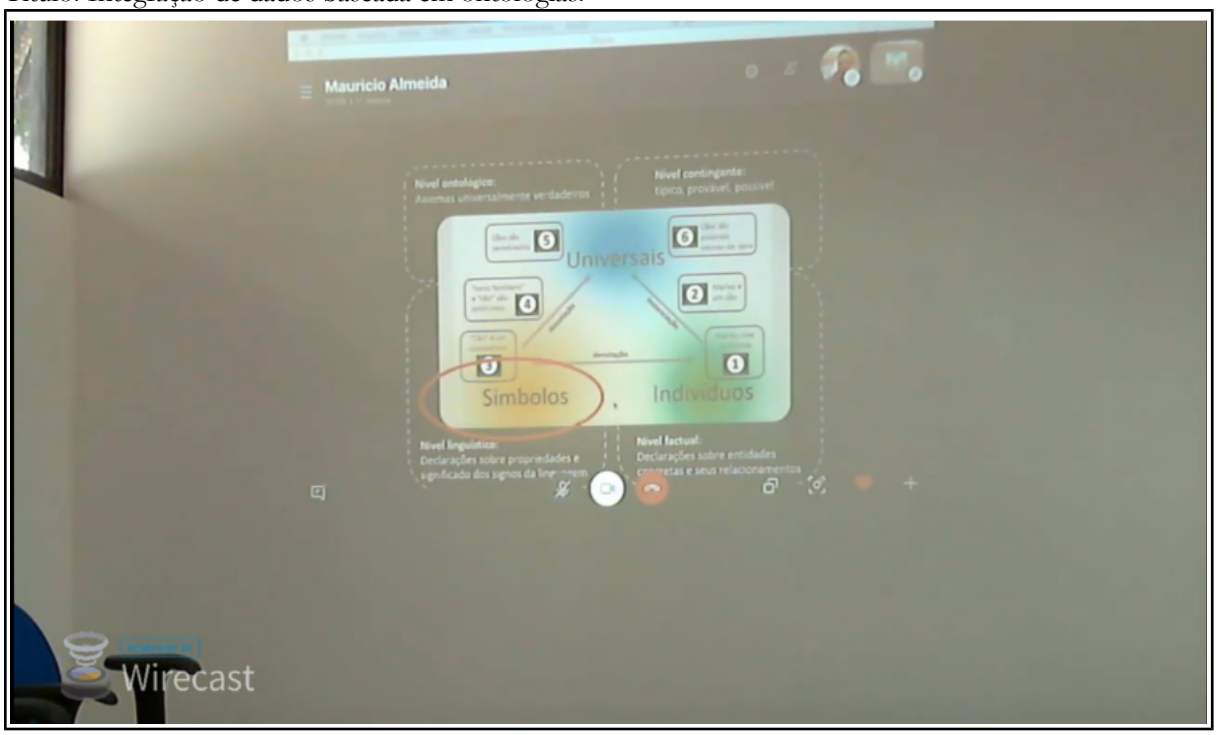
base, você tem, por exemplo, determinado osso do corpo em um hospital, no outro você tem que ter o mesmo osso em outro hospital representado basicamente da mesma forma.

Se eu for trabalhar com inferências numéricas de determinadas características do paciente, talvez a ontologia não seja capaz de representar isso tudo. Nós temos bons modelos para se fazer isso baseados em ontologias que eu posso apresentar em uma outra oportunidade porque o nosso tempo aqui é curto.

Então basicamente o que eu queria mostrar é isso, tem aqui diversas referências que foram utilizados, posso mandar depois esses slides pro Guilherme, e agradecer vocês a atenção.

Vídeo da apresentação

Título: Integração de dados baseada em ontologias.



Disponível em: http://dadosabertos.info/enhanced_publications/idt/video.php?id=29