gusttavodiias@gmail.com
lgdias@id.uff.br

Instituto de Computação

PAPER REVIEW: AN EXTENSIBLE MODELING APPROACH USING POST COORDINATED EXPRESSIONS FOR SEMANTIC PROVENANCE IN BIOMEDICAL RESEARCH

JOURNAL	AUTORES	LINK - SCOPUS	COMPILADO POR
OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"	Joshua Valdez, Michael Rueschman, Matthew Kim, Sara Arabyarmohammadi, Susan Redline, Satya S Sahoo	https://goo.gl/8uZaRY	Luiz Gustavo Dias - UFF

RESUMO

Metadados de proveniência descrevem a origem de dados para verificar e validar resultados. Reprodutibilidade de experimentos científicos vem ganhando atenção na comunidade de pesquisa em áreas como por exemplo a biomedicina. No domínio da biomedicina, foi desenvolvido ProvCaRe utilizando W3C e especificações PROV, incluindo a ontologia PROV-O. No projeto ProvCare, os autores estenderam Prov-O para criar um modelo formal de proveniência, entretanto encontraram alguns desafios associados com o desenvolvimento da ontologia, incluindo a engenharia de ontologia, redundância, e manutenção da ontologia. Em contraste com a modelagem de classes, foi proposto a “sintaxe de gramatica composicional do ProvCaRe” para modelar a ontologia sob demanda (chamado de pós-coordenação). A gramática permite reutilizar as classes de ontologias existentes e compor termos específicos de proveniência que estendem as classes e propriedades do PROV-O. A abordagem e aplicada na base do ProvCaRe, que consiste em 38 milhões de trincas de proveniências extraídas automaticamente de 384.802 artigos publicados usando um workflow de processamento de texto.

VISÃO GERAL

A reprodutibilidade é fundamental para garantir a validação de resultados de pesquisa, permitindo o avanço da ciência através do desenho rigoroso de experimentos. Portanto, a crescente adoção de técnicas de pesquisa orientadas por dados como o uso de Big Data na pesquisa biomédica e de saúde para a melhor compreensão do mecanismo de doenças e descoberta de drogas, levou a um maior foco na reprodutibilidade cientifica. Varias diretrizes e praticas foram desenvolvidas para garantir a transparência de relatórios de resultados que podem ser replicados com sucesso como por exemplo as diretrizes de Reprodutibilidade e Rigor aplicados nos institutos nacionais de saúde dos EUA (NIH). Ao alavancar os metadados de proveniência na biomedicina e saúde, os pesquisadores têm uma capacidade melhorada de colaborar, compartilhar dados e identificar as melhores praticas para pesquisas cientificas reprodutíveis.

Além da reprodutibilidade, a proveniência dos metadados também são essenciais para avaliar a qualidade dos dados e calcular o valor da confiança, e devido a sua aplicação em vários domínios, a proveniência foi modelada usando múltiplas abordagens, como por exemplo o OPM representando relação casual entre diferentes termos de proveniência.

O grupo de trabalho do W3C utilizou propriedades de abordagens de modelagem de proveniência para definir especificações do modelo PROV em 2013, que é uma representação formal de dados que utiliza a descrição baseada em OWL, chamada PROV-O (ontologia prov), e um conjunto de restrições para definir representações de origem válidas, chamado de restrições PROV. De forma geral, a ontologia W3C PROV foi desenvolvida como ontologia de referencia que pode ser estendida para modelar termos de proveniência de determinado domínio, garantindo a interoperabilidade através da utilização de um conjunto comum de classes e propriedades de ontologias.

Desta forma, os autores estenderam as especificações W3C PROV para definir uma estrutura de proveniência chamada Proveniencia Clinical and Helthcare (ProvCaRe) para apoiar a reprodutibilidade cientifica do

domínio biomédico e de saúde, que define um modelo composto por três categorias de termos de metadados de proveniência:

1. Método do estudo;
2. Ferramentas do estudo;
3. Dados do estudo;

Os autores iniciaram o desenvolvimento do ProvCaRe utilizando a pesquisa em medicina do sono como exemplo de domínio para identificar, extrair e analisar informações de proveniência associadas aos estudos, e utilizaram dados de um dos maiores repositórios do domínio o National Sleep Research Resource (NSRR). Desta forma, o objetivo foi modelar os metadados extraídos de pesquisas relacionadas a NSRR utilizando "assunto->predicado->objeto".

A ontologia W3C PROV consiste em três classes "core": prov Entity, prov Activity, e prov Agent, com nove propriedades de objeto: prov:GeneratedBy,, prov:wasDerivedFrom, prov:wasAttributedTo, prov:startedAtTime, prov:used, prov:wasInformedBy, prov:endedAtTime, prov:wasAssociatedWith e prov:actedOnBehalfOf.

MODELANDO METADADOS DE PROVENIENCIA NA ONTOLOGIA PROVCARE USANDO EXPRESSOES PÓS-COORDENADAS

A estrutura ProvCaRe modela a descrição de proveniência de um estudo cientifico que pode aumentar a capacidade de replicar o estudo por pesquisadores em outras instituições ou grupos. Os três objetivos da estrutura são: 1 criar uma ontologia de proveniência para o domínio biomédico, 2 extrair metadados de proveniência de artigos biomédicos publicados e gerar gráficos de proveniência para analise, e 3 desenvolver uma base de conhecimento de proveniência para usuários pesquisarem e identificarem estudos de pesquisa que possam ser replicados para validar resultados importantes ou projetar novos estudos experimentais.

Conforme descrito, os três componentes da estrutura ProvCaRe representam três tipos de informações essenciais de proveniência correspondentes ao método usado para conduzir o estudo (método), dados utilizados (dados do estudo), e detalhes dos instrumentos utilizados no estudo (instrumento de estudo). As informações de proveniência destes três componentes podem ser modeladas em detalhes, o que e importante para capturar com precisão as informações contextuais necessárias para replicação dos estudos anteriores.

METADADOS DE PROVENIENCIA EM PESQUISAS DE MEDICINA DO SONO

O NSRR é o maior repositório de dados de medicina do sono de pesquisa disponível, e inclui dados de pesquisa que são representativos em uma ampla variedade de tópicos, portanto é indispensável que a estrutura do ProvCaRe seja escalável. Estudos de biomedicina são frequentemente definidos em termos do conhecido modelo de população, intervenção, comparação, resultado e tempo (PICO(T)) (O = outcome = resultados), para representar diferentes aspectos de um estudo de pesquisa. Entretanto o modelo PICO(T) não inclui muitos termos de proveniência critica como por exemplo, não representa os termos de proveniência correspondentes ao método de análise de dados (modelo estatístico usado para derivar os resultados do estudo), instrumentos usados para registrar dados (tipo de instrumento de pressão arterial usado no exemplo da pesquisa). Para resolver esse problema os autores estenderam a ontologia W3C PROV no projeto ProvCaRe para modelar informações de proveniência correspondentes aos três aspectos de um estudo cientifico, a saber, ferramentas de estudo, método e dados.

SINTAXE DE GRAMATICA COMPOSICIONAL NA ONTOLOGIA PROVCARE

Os autores estenderam e adaptaram a sintaxe de gramatica composicional pos-coordenacao definida, para criar expressões SNOMED CT (terminologia clinica utilizada em mais de 50 paises), para a estrutura ProvCaRe usando classes e propriedades definidas na ontologia ProvCaRe e ontologias biomédicas existentes.

Uma expressão pós-coordenada do ProvCaRe consiste em uma única classe de ontologia, que e o conceito de origem central da expressão e um conjunto de propriedades, bem como seus valores, que qualificam o conceito central. As propriedades e os valores associados podem ser definidos no ProvCaRe , ou os valores podem ser valores literais de RDF.

Quatro categorias de expressões pós-coordenadas podem ser compostas usando a gramatica ProvCaRe:

- Expressao multi-conceito: duas ou mais classes de ontologia pode ser combinadas utilizando o sinal "+" para formar um novo conceito, que e interpretado como uma subclasse das duas classes originais. Por exemplo, um método de analise de dados pode envolver: provcare: CorrelationAnalysis e procare: CovarianceAnalisis, que pode ser representado usando a expressão: |CorrelationAnalysis|+|CovarianceAnalysis|.
- Conceito com restrições definidas sobre propriedades: um termo de proveniência pode ser refinado usando restrições adicionais definidas sobre uma restrição ProvCare ou outra prorpeidade da ontologia. Por exemplo, e importante registrar a proveniência da pressão arterial de um paciente do estudo de pesquisa em termos do procedimento. Esta informação de proveniência pode ser modelada usando uma expressão que combina a ontologia ProvCare e termos do SNOMED CT: |Electrocardiograph|: |hadDataCollectionMethod| = |12 lead ECG| (ECG has SNOMED CT ID: C0180600 and 12 lead ECG has SNOMED CT ID: C0430456). Um conceito central também pode incluir varias restrições definidas usando varias propriedades. Por exemplo: |Electroencephalogram|: |hadStudyInstrument| = |Scalp electrode cap|, |hadLocation| = hospital|
- Conceitos com restrições definidas sobre grupos de propriedades: A gramatica de composição ProvCaRe permite o agrupamento de múltiplas propriedades em uma subunidade, para reduzir a ambiguidade em relação a ordenação das restrições usando uma abordagem que e similar a gramatica composicional do SNOMED CT. Por exemplo, dois ECG podem ter sido realizados em dois locais diferentes, que podem ser representados usando a expressão: | Electrocardiograph |: {| hadDataCollectionMethod | = | 12 ECG de chumbo |, | hadLocation | = | hospital |}, {| hadDataCollectionMethod | = | 12 ECG de chumbo |, | hadLocation | = |Home |}. As chaves colocam duas ou mais propriedades juntas para permitir aos usuários e ferramentas de software, analisar corretamente a ordenação das restrições de uma expressão.
- Conceitos com restrições aninhadas: Como discutido anteriormente, as expressões pos-coordenadas ProvCare usam uma definição recursiva, que permite que o valor na estrutura da tripla seja outra expressão. Por exemplo, um estudo que usa duas técnicas para analise de dados estatísticos podem ser representadas usando a seguinte expressão: | ResearchStudy |: | hadDataAnalysisMethod | = (| CorrelationAnalysis | + | CovarianceAnalysis |). A estrutura aninhada também pode ser contruida usando varias propriedades, por exemplo: | ResearchStudy |: | hadDataAnalysisMethod | = (| StatisticalMethod |: | hadStatisticalMeasure | = | CentralTendencyMeasure |).

RESULTADOS

Foi demonstrado o uso pratico de expressões pós-coordenadas na ontologia ProvCaRe usando um exemplo de estudo de pesquisa publicado por O'Connor et al, e a efetividade da ontologia ProvCaRe na extração de 38 milhoes de triplicas de procedência de 384.802 artigos publicados.

Referente a expressões pós-coordenadas: O estudo de pesquisa de O'Connor et al, e classificado como estudo observacional (modelado como subclasse de Design na ontologia ProvCaRe). Usando a gramatica de composição do ProvCaRe, modelar as informações de proveniência relacionadas a analise de dados usando classes de ontologias e propriedades modeladas na ontologia ProvCaRe e nas ontologias biomédicas existentes. Para exemplo, a expressão pos-coordenada | ResearchStudy |: | hadDataAnalysisMethod | = | regressão multivariada |, | hadSoftwareTool | = | SAS |, usa termos de regressão multivariada e SAS modelados na ontologia bilingue da doença de Alzheimer e doenças relacionadas (ONTOAD) e software de ontologia (CWO). Desta forma, a população selecionada para a pesquisa pode ser caracterizada usando a seguinte expressão: | ResearchStudy |: {| hadStudyConstraint | = (| StudyExclusionCriterion |: | hadPrescription | Medicação anti-hipertensiva |)}. Esta expressão apresenta metadados importantes de proveniência descrevendo as restrições usadas para identificar participantes para a pesquisa, e é essencial para outros pesquisadores que pretendem replicar o estudo.

É importante notar que as expressões pós-coordenadas relacionadas a procedência precisam ser criadas frequentemente por especialistas de domínio com pouca ou nenhuma experiencia em engenharia de ontologia. Portanto, o desenvolvimento de um vocabulário de interface visual pode ajudar significativamente na criação de uma expressão pós-coordenada válida. A composição do ProvCare suporta o desenvolvimento de um modelo de entrada do usuário com base em formulário que usa valores da propriedade como widgets e as classes de ontologia correspondentes como valores.

Referente a criação da base de conhecimento: A extração de dados estruturados a partir de texto e um desafio significativo e este tem sido um foco de extensa pesquisa em ciência da computação usando aprendizado de maquina, bem como técnicas baseadas em regras. O uso de técnicas da web semântica especialmente usando

ontologias como modelo de conhecimento de referencia tem sido uma abordagem eficaz no processamento de linguagem natural. No entanto, os autores não tinham conhecimento de nenhum trabalho anterior que usasse ontologias para extrair metadados de proveniência de texto não estruturado. Para extrair e analisar as informações de proveniência de estudos biomédicos, eles desenvolveram um workflow de processamento de linguagem natural usando a ontologia ProvCaRe. Usando o workflow habilitado para a ontologia, foi possível processar e extrair informações de proveniência de quase 400.000 trabalhos, descrevendo estudos.

A base de conhecimento suporta consulta usando duas abordagens: 1 orientada por hipóteses, para procurar pesquisas anteriores correspondentes a uma determinada hipótese e ver os metadados de proveniência associados a cada um desses estudos para fins científicos; 2 usar as informações de proveniência de estudos anteriores para projetar novos experimentos com protocolos rigorosos com a finalidade de garantir relatórios transparentes bem como suporte a reprodutibilidade.

ANÁLISE SEGUNDO O LEITOR

1. Qual tipo de proveniência abordada no trabalho?

   Essa informação não é explicita no texto, entretanto os autores trabalham com metadados de proveniência referentes aos passos da pesquisa, dessa forma, assume-se que a proveniência abordada no estudo é a prospectiva: "The ProvCaRe framework defines a reference model consisting of three categories of provenance metadata terms that we have identified as necessary for scientific reproducibility in biomedical research: 1. Study Method: The design of the research study in terms of study design, sampling, randomization technique and interventions (in experiments), data collection approach, and data analysis techniques (e.g., statistical models) are examples of provenance metadata describing Study Method; 2. Study Tools: The different instruments and their parameter values that are used to record and analyze data in a research study is the second essential component of the ProvCaRe framework. For example, the strength of the magnet used in a Magnetic Resonance Imaging (MRI) instrument is important provenance information that will allow other researchers to use an equivalent MRI machine to replicate the findings of the original experiment; and 3. Study Data: The provenance metadata describing the contextual information about the data elements used in a scientific experiment, for example drug information, demography information of participants, and timestamp values, are necessary to allow other researchers to replicate a given experiment."

2. Qual tipo de ontologia utilizada no trabalho?
   É feita uma extensão da ontologia PROV-O para o domínio da biomedicina

3. Qual a principal vantagem em se utilizar ontologia no contexto da pesquisa?
   A principal vantagem do uso de ontologias percebida no artigo é a disseminação e padronização da informação, possibilitando geração de conhecimento, uma vez que os termos do domínio que compõem a ontologia, são definidos de forma formal.

4. Questões de granularidade são abordadas no artigo?
   Não. A granularidade não é abordada no artigo.