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

Instituto de Computação

PAPER REVIEW: A GENERAL PURPOSE TOOL-SET FOR REPRESENTING DATA RELATIONSHIP: CONVERTING DATA INTO KNOWLEDGE

JOURNAL	AUTORES	LINK - SCOPUS	COMPILADO POR
Scientific Data Summit (NYSDS), 2016 New York. IEEE, 2016.	Joshua Stillerman, Thomas Fredian, Martin Greenwald, John Wright	https://goo.gl/m1EFFw	Luiz Gustavo Dias - UFF

RESUMO

Metadados soa necessários para entender experimentos modernos que geralmente possuem grandes e complexas bases de dados. Sistemas para armazenar e gerenciar esses metadados melhoram com o tempo, mas a maioria de sistemas para esta finalidade muitas vezes são representados por um aplicativo de domínio especifico. Desta forma os autores propõem o desenvolvimento de um conjunto de ferramentas para armazenar, gerenciar e recuperar o relacionamento de dados e metadados. Essas ferramentas são aplicáveis em uma ampla gama de domínios. Ferramentas para gerenciamento de dados geralmente representam pelo menos um paradigma de relacionamento através de metadados implícitos ou explícitos. A adição de metadados permite que dados sejam pesquisados e entendidos por parte dos usuários por períodos mais longos. Desta forma, os pesquisadores se tornam menos dependentes de outros cientistas envolvidos na execução do experimento. Na comunidade de pesquisa de fusão magnética, o sistema MDSplus e amplamente utilizado para registrar dados brutos e processados de experimentos. Usuarios criam uma arvore de relacionamento para cada instancia do experimento. No entanto, a arvore MDSplus e apenas possível para organização dos registros, e outros aplicativos relacionam o experimento.

MPO e um sistema construído para registrar informações de proveniência de dados sobre resultados computados. Permite aos usuários gravar entradas e saídas de cada etapa do workflow. Desta forma os resultados geram gráficos de proveniência e podem ser anotados, agrupados, pesquisados e filtrados. Isso fornece uma ferramenta poderosa para registrar, entender e localizar resultados computados. No entanto, isso pode ser entendido como uma relação de dados mais especifica, que pode ser interpretado como uma instancia de algo mais geral. Com base em conceitos gerais, os autores propõem um sistema que possa ser usado para representar todos os tipos de relacionamentos de dados como gráficos matemáticos. Assim como o MDSplus e MPO foram generalizações de gerenciamento de dados, este novo sistema generalizara armazenamento, localização e recuperação das relações entre os dados. Dados armazenados seriam referenciados por URIs permitindo que o sistema seja agnóstico para representações de dados subjacentes. Os usuários podem em seguida, percorrer os gráficos gerados. O sistema permitirá que os usuários construam uma coleção de gráficos descrevendo qualquer ou toda relação entre os itens, localizando dados de interesse, consultando outros gráficos e navegar entre eles.

VISÃO GERAL

A coleta de dados evoluiu da forma manual para uma abordagem automática, desta forma e relativamente fácil adquirir gigabytes de dados. Ao mesmo tempo, o custo de armazenamento para esses dados tem diminuído constantemente. Isso possibilita o registro de conjuntos de dados extremamente grandes. A facilidade com que os dados podem ser coletados torna imperativo que os cientistas envolvidos organizem os dados. Cientistas costumavam cuidadosamente arquivar ou colar em cadernos de laboratórios gráficos e fotos, fazendo com que fosse razoavelmente fácil encontrar e entender posteriormente, assumindo que a cronologia era um princípio de organização suficiente.

O crescimento em tamanho e complexidade de conjuntos de dados científicos exacerba os problemas associados a descoberta de dados e compreensão. Projetos grandes são produtos de colaborações com equipes

grandes e geralmente distribuídas geograficamente. Pesquisas em física, ciências da terra, ciências da vida, genética, ciências sociais, todos compartilham dessas preocupações.

ABORDAGEM

- **Generalizar as representações de relacionamentos de dados:** Um esquema geral de metadados precisa ser capaz de se referir aos dados armazenados. Este e o caso dentro de um único projeto, pois há quase sempre uma variedade de tipos de dados que o sistema deve representar. Para o sistema ser aplicável a vários projetos em diferentes domínios de investigação isso e ainda mais imperativo, já que eles improvavelmente compartilham mecanismos subjacentes de armazenamento de dados. Recurso uniforme de identificação (URI) fornecem um mecanismo para homogeneizar recursos heterogêneos. Um URI e uma sequencia de caracteres usada para identificar um recurso. Essa identificação permite a interação com representações do recurso em uma rede utilizando protocolos específicos. O sistema precisa estar ciente da complexidade e granularidade dos objetos de dados a que se refere e as necessidades de interfaces uteis e voltadas para o usuário.
- **Relacionamentos são gráficos:** Relacionamentos entre dados armazenados são representados como gráficos. Isso e, coleções de vértices (nós) e arestas (conexões). Existem relativamente poucas classes de gráficos que o sistema precisara representar. As coleções desses meta-esquemas formam os esquemas dos dados de relacionamento para cada domínio de aplicação. Ferramentas podem ser desenvolvidas para operar cada tipo de gráfico. Essas ferramentas podem ser aplicadas a todos os gráficos desse tio no esquema especifico do domínio, desta forma os autores planejam se apoiar em arvores hierárquicas, ordenadas e desordenadas, linhas do tempo, gráficos acíclicos dirigidos. Essa lista será expandida como necessidade especificas de domínio.
- **Esquema como dados:** Os esquemas, isto é, coleção de tipos de gráficos, para um determinado domínio será representado como dados uniformemente acessíveis, em vez de código de aplicativo ou estruturas de dados disjuntas. Isso permite que os administradores e usuários adicionem novos relacionamentos que podem não ser explorados pelo usuário comum. É planejado também implementar uma interface web para interagir com estes esquemas.
- **Instancias como dados:** Os dados de relacionamento podem ser divididos em três categorias baseadas em sua taxa de evolução, embora possam compartilha representação e implementação. Cada uso do sistema terá coleções de metadados descritivos estáticos que descreverão o projeto como um todo. Os nos desses gráficos podem conter URIs para recuperar registros detalhados sobre o projeto. A próxima classe de metadados e aplicada a mudança de quantidades. Novamente eles podem ou não conter URIs para recuperar objetos. A classe final de metadados descreve os registros de dados armazenados do experimento. Os nos desses gráficos contem URIs para instancias de dados recuperáveis. Como acontece com esquemas, será fornecido uma interface web para ser preenchida e explorar os metadados descritos. Depois de navegar para qualquer referência de dados, os usuários poderão perguntar: de que outros gráficos esse objeto e membro? Isto é, que outros dados são relacionados a esse item? Quais são seus vizinhos neste gráfico? Fornecendo um ambiente rico para documentar e explorar os resultados armazenados.

EXEMPLOS

- **Biblioteca:** Uma biblioteca fornece uma analogia simples que ilustra as principais características do sistema. Uma biblioteca contem uma coleção de recurso, banco de dados etc. Usando um catalogo, os usuários podem rastrear um recurso especifico. E se pudessem navegar facilmente por itens com diferentes critérios de adjacência? E se esses critérios fossem extensíveis e anotáveis? E se os novos critérios e anotações fossem acessíveis a todos os clientes da biblioteca?
- **Pesquisa de fusão magnética:** Grande parte da pesquisa mundial de energia de fusão magnética usa o sistema MDSplus para armazenar e organizar resultados adquiridos e computados. A hierarquia, ou arvore, permite aos usuários criar associações entre itens de dados armazenados. Por exemplo, um diagnostico de temperatura pode ter uma estrutura de arvore com descrições textuais e resultados finais computados no topo, e os detalhes dos dados adquiridos e cálculos nos níveis inferiores da arvore. Usuários familiarizados com o layout geral da arvore podem navegar e pesquisar para encontrar as medidas em que se tem interesse. Uma vez encontrados, os nos associados próximos a eles fornecem um contexto.

- **Projeto de ontologia de proveniência de metadados:**O sistema de ontologia de proveniência de metadados (MPO) controi gráficos de proveniência de workflows. Esses gráficos acíclicos dirigidos podem ser usados para localizar e compreender as origens dos resultados computados, mostrando quais quantidades adquiridas e calculadas foram utilizadas no pipeline computacional, os gráficos podem ser agrupados em coleções, anotados como comentários e ampliados como metadados estruturados. A navegação nestes gráficos fornece o contexto necessário para entender significado e proveniência dos resultados armazenados.

CONCLUSÃO

As relações entre os dados são o que dá significado a eles. Preservando e documentando essas relações, permite que esses dados sejam usados por grupos mais amplos de usuários em períodos mais longos de tempo. Ao longo do tempo, a coleta de dados, organização e recuperação sofreu generalização sucessiva. Como isso ocorreu, tornou-se mais fácil para os usuários explorar seus dados. Os autores propõem assim, criar um sistema para representar relacionamentos entre dados armazenados em um modo gera, extensível e orientado por dados. O sistema facilitara a exploração e a descoberta de atividades sobre armazenamento de dados de pesquisas. Os dados serão representados pela URI, independentemente da sua origem ou mecanismo de armazenamento, o que permitira que o sistema opere em uma ampla variedade de dados armazenados e em domínios de pesquisa diferentes.

ANÁLISE SEGUNDO O LEITOR

1. Qual tipo de proveniência abordada no trabalho?
   Essa informação não é explicita no texto, entretanto os autores do trabalho propõem uma ferramenta para captura de proveniência retrospectiva, aplicável em experimentos baseados em workflow, independente de domínio.

2. Qual tipo de ontologia utilizada no trabalho?
   Não existem informações sobre o tipo de ontologia utilizada, entretanto percebe-se que foi utilizada uma ontologia de tarefa, tomando como base a definição de Isotani e Bittencourt (2015) "a ontologia de tarefa representa os processos e atividades para resolver um determinado problema abstraindo o contexto do domínio"

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?
   A granularidade não é explorada a fundo, entretanto a partir do trecho " The system needs to be cognizant of the complexity and granularity of the data objects it refers to and the needs of usable, user-facing interfaces." Assume-se que a ferramenta será apta a trabalhar com granularidades de diferentes tipos.