Uma nova abordagem para consulta a dados de proveniência

Patricia Vieira Camara Costa[1], Vanessa Braganholo[1,2]

[1]Programa de Pós-Graduação em Informática (PPGI), UFRJ
[2]Instituto de Computação, UFF

pattyvieiracosta@gmail.com, vanessa@ic.uff.br

Abstract. *One of the most important features of scientific experiments is the reproducibility. To achieve reproducibility, scientists have been automating their experiments using scientific workflows which allows provenance data to be automatically captured. Provenance describes the origin of a piece of data (program and parameters used to generate it, the user who ran the program, date and time, among others). However, there is no point in collecting this data is it cannot be queried satisfactorily. The problem is that to query this data is not always an easy task, since it requires scientists' to have expertise on specific query languages, which they usually do not have. This work proposes a method for assembling provenance queries aiming at reducing the effort of scientists.*

1. Introdução

Ao longo dos anos, os experimentos científicos vêm se tornando cada vez mais complexos, gerando um volume de dados cada vez maior (Mattoso *et al.*, 2008). Diante desse cenário surge a necessidade de alternativas que venham a auxiliar os cientistas na construção desses experimentos. Uma alternativa que vem sendo cada vez mais adotada é o uso de *workflows* científicos. Estes *workflows* são gerenciados por Sistemas de Gerência de *Workflow* Científico (SGWfC). Os SGWfC têm como propósito apoiar as várias etapas do processo de construção do experimento, orquestrar sua execução, coletar dados de proveniência, além de permitir diversas análises (Gil *et al.*, 2007). Dessa forma é possível reproduzir o experimento tal como foi projetado originalmente.

Dados de proveniência ajudam a determinar a história do dado produzido em um experimento a partir de suas fontes originais (Freire *et al.*, 2008). O gerenciamento de proveniência auxilia os cientistas a interpretar e compreender os resultados que são obtidos nos experimentos realizados a partir de um SGWfC. É possível, por exemplo, saber qual foi a sequencia de fatos que deu origem a determinado artefato do *workflow*.

Porém, nota-se que poucos sistemas tratam da questão de consulta a esses dados. Essa é uma tarefa que vem exigindo que o usuário tenha conhecimento sobre como escrever consultas em linguagens relativamente complexas. No entanto, nem sempre o cientista possui tal conhecimento. Ao mesmo tempo, essas consultas são de extrema importância para o cientista, pois estão ligadas a questões críticas como processo de patentes e de descoberta científica, entre outros. Para mitigar este problema, a proposta desse trabalho é criar um método que seja capaz de auxiliar o cientista na geração de

consultas a dados de proveniência sem exigir dele conhecimentos especiais para a criação dessas consultas.

Esse artigo está organizado em três outras seções. A seção 2 apresenta o conceito de proveniência de dados em *workflows* científicos, mostrando a questão de captura, armazenamento e consulta aos dados de proveniência. Na seção 3 é apresentado um método para montagem de consultas aos dados de proveniência anteriormente coletados e armazenados. Finalmente, a seção 4 apresenta as considerações finais destacando as contribuições e possíveis trabalhos futuros.

2. Trabalhos relacionados

A proveniência auxilia na coleta de informações históricas dos dados que são manuseados em um *workflow*. Freire *et al*. (2008) apresenta três níveis de captura de proveniência: Sistema Operacional (SO), no qual dados de proveniência são capturados em grão fino a partir de eventos do SO; SGWfC que é fortemente acoplado ao sistema de origem do *workflow* e captura os dados em um grão mais grosso; Atividade, onde cada atividade do *workflow* torna-se responsável por capturar seus próprios dados de proveniência. Isso mantém a independência em relação ao SGWfC, mas ao mesmo tempo exige que cada atividade seja devidamente adaptada para que a captura possa ocorrer.

Na literatura diversos SGWfC tratam da questão de proveniência de dados em *workflows* científicos. O *Pegasus* (*Planning for Execution in Grids*), por exemplo, é um sistema gerenciador de *workflow* que faz uso do VDL (*Virtual Data Language*), linguagem criada para a captura e consulta de dados de proveniência. Ele armazena seus dados em um Sistema Gerenciador de Banco de Dados (SGBD) (Foster *et al.*, 2002). O sistema captura dados ligados ao momento de desenvolvimento do *workflow*, através da linguagem de consulta ***S****PARQL* ***P****rotocol* ***a****nd* ***R****DF Query Language* (SPARQL) e dados ligados ao momento de execução do *workflow*, usando a linguagem SQL.

Já o *VisTrails* é um SGWfC que armazena seus dados, dentre eles os de proveniência, em formato *eXtensible Markup Language* (XML) ou em um SGBD (*MySQL* ou DB2). Ele possui três componentes principais: *Builder*, que monta, gerencia e executa o *workflow*; *VisTrails Server*, responsável pelo armazenamento dos *workflows* construídos; *VisTrails SpreadSheet*, responsável pela visualização dos resultados do workflow (Bavoil *et. al.*, 2005). Esse sistema possui uma interface de consulta que permite que dados sejam recuperados através de consultas *XPath* ou *XQuery* (Bavoil *et al.*, 2005). Define também uma linguagem própria, *VisTrails Query Language*, que permite consultas a versões de *workflow* e a recuperação de informações em *log* de execução.

O Kepler é um SGWfC orientado à atores, ou seja, possui os papéis de diretor, responsável por controlar a execução do *workflow* e ator, responsável pela modelagem das tarefas (Menezes, 2008). Com relação à parte de consulta, o Kepler trabalha com a interface de consulta *Query Actor*, que exige do usuário conhecimento da linguagem SQL para montar as consultas.

Anand *et al.* (2010) propõem em seu trabalho que consultas de proveniência sejam feitas através de um navegador de proveniência. Esse navegador faz uso da

linguagem de consulta *Query Language for Provenance* (QLP) que tem por objetivo expressar de maneira clara as consultas complexas de proveniência. Permite aos usuários navegar pelo grafo de proveniência sob diferentes perspectivas gráficas, além de manter o histórico das consultas previamente montadas permitindo dessa maneira que o usuário volte a navegar em resultados de consultas anteriores. O *browser* é integrado ao SGWfC Kepler.

Já Marinho *et al*. (2009) desenvolveram o *ProvManager*, um mecanismo de captura de proveniência em *workflows* científicos para ambientes distribuídos e que independe de um SGWfC para sua execução. Para isso foi adotado um mecanismo de captura de proveniência no nível de atividade. O repositório de proveniência do *ProvManager* é formado por um banco de dados Prolog. O *ProvManager* trabalha com agentes inteligentes através do *Provenance Viewer Agent*. Tem como vantagem o fato de não estar preso à tecnologia de execução do *workflow*.

O *ProvManager* já mostra uma preocupação na literatura com mecanismos que independam do SGWfC, deixando o usuário livre para escolher o que mais lhe convém. No entanto, a linguagem de consulta adotada é a Prolog. Consultas Prolog pré-montadas são oferecidas aos usuários de tal forma que ele precise apenas mudar os parâmetros das consultas para obter os resultados desejados. É possível também que o usuário monte suas próprias consultas (Marinho *et. al.*, 2010). Em ambos os casos, o usuário precisa conhecer Prolog para utilizar a interface de consulta. Ainda há, portanto, espaço para aprimorar os mecanismos de consulta a dados de proveniência, de modo a não exigir do cientista nenhum conhecimento sobre linguagens de consulta. A próxima seção apresenta a abordagem proposta para a construção de tais consultas.

3. Um mecanismo de consulta para dados de proveniência

Esse trabalho tem como proposta elaborar um método que seja capaz de reduzir o esforço dos cientistas no momento da construção de consultas a dados de proveniência em *workflows* científicos. Com isso, será retirada do usuário a tarefa de construir consultas complexas, em linguagens muitas vezes desconhecidas por eles.

Para a realização desse trabalho optou-se pelo uso de dados oriundos do *ProvManager*, por ele ser um sistema de captura de dados de proveniência que independe do SGWfC (Marinho *et al*., 2009). Isso faz com que a consulta seja mais ampla e não fique dependente de uma única ferramenta, pois no geral os SGWfC's trabalham com uma linguagem própria. Toda proveniência coletada pelo *ProvManager* está armazenada em fatos Prolog, o que exige do usuário um conhecimento prévio dessa linguagem. Ao mesmo tempo, o Prolog permite trabalhar com inferências e, com isso, as consultas geram respostas mais completas. Portanto, a finalidade é eliminar a necessidade do cientista de conhecer a linguagem Prolog para que as consultas sejam montadas e seus resultados analisados.

A abordagem proposta prevê que o cientista defina sua consulta em alto nível, utilizando um formulário para escolha dos critérios da consulta, por exemplo. De posse da especificação de consulta em alto nível, o sistema aqui proposto transforma os dados recebidos do formulário em uma consulta do tipo Prolog. Tendo em mãos a consulta Prolog construída, a terceira etapa consiste em executá-la. Para tal faz-se necessário o uso de uma máquina Prolog. Uma possível alternativa para ser empregada é o *SWI-*

Prolog, que é uma implementação em código aberto da linguagem Prolog (Wielemaker, 1996). No entanto, o resultado obtido nessa fase é um resultado de consulta Prolog. Dessa forma, a quarta e ultima etapa é responsável por transformar os dados recebidos, em Prolog, de tal forma que os mesmos sejam visualizados de maneira clara e objetiva pelo usuário. A Figura 1 mostra cada uma dessas etapas.

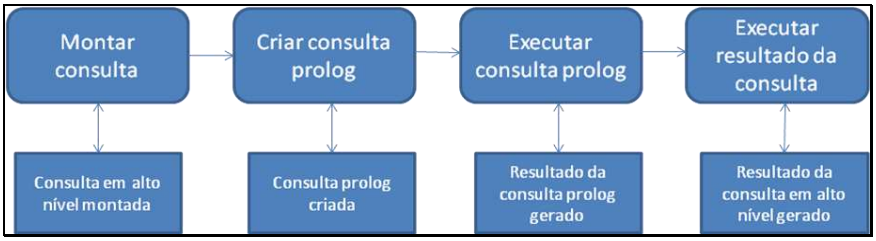

Figura 1 - Passo a passo para construção das consultas

Inicialmente a ideia é criar um formulário em alto nível para que os cientistas possam selecionar os dados necessários à montagem da consulta. Contudo, outras estratégias para a definição da consulta serão pesquisadas a fim de encontrar uma abordagem melhor adaptável a esse contexto. Uma das ideias para tal tarefa é fazer uso de recursos presentes em ferramentas do tipo OLAP, como técnicas de arrastar e soltar que auxiliem na montagem da consulta.

Os sistemas OLAP oferecem também uma resposta rápida e consistente às consultas montadas, por mais complexas que elas possam ser (Figueiredo, 1998). Por outro lado, já existem na literatura diversas ferramentas que tratam da questão de visualização de maneira satisfatória. Estudos foram realizados a respeito da união de ferramentas OLAP com o repositório Prolog usado nesse trabalho, pois essas ferramentas usam tuplas e o repositório Prolog trabalha com fatos. Faz-se necessário a criação de uma fase de mapeamento entre esses dados de tal forma a não haver conflitos.

Ao mesmo tempo o Prolog auxilia bastante no momento de realização de consultas, pois uma série de informações de procedência pode ser obtida através de inferências feitas a partir dos dados de proveniência do *workflow*. Dessa forma podem ser geradas respostas mais completas, que vão além de afirmativas ou negativas para as questões abordadas nas consultas.

Além da transformação do resultado para uma linguagem em alto nível a ideia desse trabalho é fazer uso de técnicas de visualização de informação visando uma melhor apresentação desses dados. Dessa forma, uma possível solução a ser incorporada nesse trabalho seria o uso de técnicas de visualização de informação por categorias como, por exemplo, a navegação facetada. De acordo com Barbosa (1972) a classificação através de facetas tem como finalidade sintetizar conceitos por mais que os mesmos sejam diferentes. Nesse caso, as categorias e hierarquias são construídas manualmente e, com isso, faz-se necessário um conhecimento prévio do usuário a respeito do domínio em que o mesmo vai trabalhar.

Um exemplo prático dessa técnica são as listas elásticas (*elastic lists)*(Stefaner e Muller, 2007). Os itens da lista são ordenados e o tamanho de um determinado item dessa lista indica a proporção de itens da consulta que estão associadas com aquele metadado em questão. Tal técnica facilita a montagem da consulta, foco principal desse

trabalho, além de apresentar os resultados de maneira satisfatória. Além disso, é possível dar continuidade ao uso do Prolog como executor das consultas, o que não seria possível caso fosse usado um ambiente OLAP, pois o mesmo exige que o banco de dados Prolog seja adaptado ou traduzido para tuplas. Dessa forma, esse trabalho optou por utilizar listas elásticas para resolução do problema proposto.

A Figura 2 mostra um protótipo não funcional do sistema. Foram criadas cinco facetas dentre as quais os cientistas poderão escolher parâmetros para construção da consulta. A primeira aba, experimento, lista todos os experimentos presentes na aplicação. Na segunda aba é possível escolher quais os *workflows* fazem parte dos experimentos selecionados anteriormente. A aba atividade fornece a lista de atividades dos *workflows* selecionados e aba parâmetros mostra os parâmetros de cada uma das atividades anteriormente indicadas. Em seguida é possível salvar a consulta, antes que ela seja executada, ou então simplesmente executá-la sem que a mesma seja salva na lista de consultas do usuário.

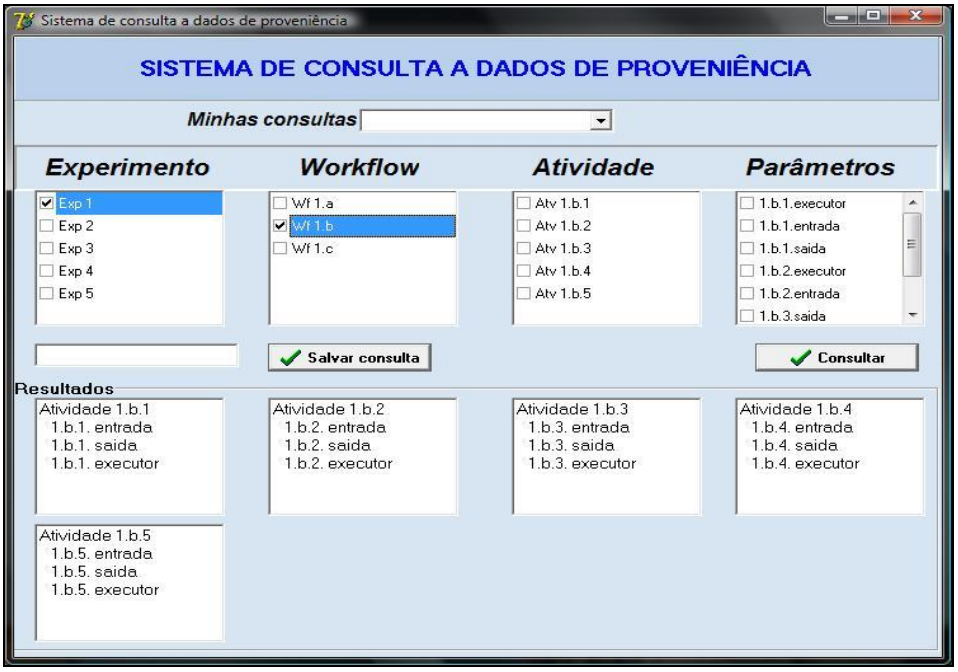

Figura 2 – Protótipo do sistema de consulta

4. Conclusão

Este artigo apresenta uma visão geral a respeito da questão da gerência de dados de proveniência em *workflows* científicos, destacando como um ponto importante a questão de consulta a esses dados. Percebe-se que uma das coisas mais importantes para os cientistas é a análise dos dados que são coletados nas experiências e, para isso, faz-se necessário a existência de um sistema de consulta a esses dados. Esse sistema tem como finalidade ajudar no trabalho de interpretação dos resultados obtidos nas experiências científicas.

Diante desse cenário, é proposta a criação de um método para construção de consultas de tal forma que não seja necessário que o usuário tenha conhecimentos prévios e complexos para a realização dessa tarefa, além da aplicação de técnicas que

permitem ao cientista uma melhor análise e visualização dos resultados que são obtidos nos experimentos. Espera-se alcançar tal propósito através da criação de um sistema que usará como interface a técnica de listas elásticas.

A escrita dos três primeiros capítulos da tese já foi feita. Atualmente esse trabalho encontra-se na fase de implementação do sistema. Os próximos passos consistem no término da implementação do sistema que dará suporte ao método e a aplicação em algum ambiente de pesquisa científica que venha a comprovar, ou não, a hipótese levantada de que através desse método de montagem de consulta apresentado, o trabalho do cientista venha a ser facilitado.

Referências

Anand, M. K, Bowers, S., Altintas, I., Ludascher, B. (2010) "Approaches for Exploring and Querying Scientific Workflow Provenance Graphs". In: International Provenance and Annotation Workshop (IPAW), Troy, NY, EUA.

Barbosa, A. P. "Classificações facetadas". Ciência da informação, Rio de Janeiro, v. 1, n. 2, p. 73-81. (1972).

Bavoil, L., Callahan, S., Crossno, P., Freire, J., Scheidegger, C, Silva, C. and Vo, H. (2005) "Vistrails: Enabling interactive multipleview visualizations" In: IEEE Visualization, p. 135-142.

Figueiredo, A. M. C. M. (1998) "Molap x Rolap: Embate de Tecnologias para Data Warehouse, Developers' Magazine". v. 2, n. 18, p. 24-25.

Freire, J., Koop, D., Santos, E., Silva. and C. T. (2008) "Provenance for Computational Tasks: A Survey". Computing in Science and Engineering, v. 10, n. 3, p. 11-21.

Foster, I., Voeckler, J., Wilde, M. and Zhao, Y. (2002) "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation". In: Conference on Scientific and Statistical Database Management.

Gil, Y., Deelman, E., Mark H. Ellisman, Fahringer, T., Geoffrey, F., Dennis G., Carole A., Miron, L., Luc M. and Jim M. (2007). "Examining the Challenges of Scientific Workflows". IEEE Computer, v. 40(12), p. 24-32.

Marinho, A., Murta, L., Werner, C., Braganholo, V., Ogasawara, E., Cruz, S.M.S., Mattoso, M., (2010) "Integrating Provenance Data from Distributed Workflow Systems with ProvManager." In Proceedings of the International Provenance and Annotation Workshop (IPAW).

Marinho, A., Murta, L. G. P., Werner, C., Braganholo, V., Cruz, S. S. and Mattoso, M. L. Q. (2009), "A Strategy for Provenance Gathering in Distributed Scientific Workflows", In: SWF, Los Angeles, California.

Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Murta, L., (2008), "Gerenciando Experimentos Científicos em Larga Escala", In: SEMISH - CSBC, Belém, Pará - Brasil.

Menezes, J. G. M. (2008) "Gerência Distribuída de Dados em *Workflow* de Bioinformática." Dissertação de mestrado do Instituto Militar de Engenharia, Rio de Janeiro

Stefaner, M. and M¨uller, B. "Elastic lists for facet browsers". (2007). In: International Conference on Database and Expert Systems Applications,Washington, DC, USA.

Wielemaker, J. (1996). "SWI-Prolog Reference Manual. University of Amsterdam". Disponível em <http://www.swi-prolog.org/>, Acessado em Abril, 14, 2010.