O Uso de Redes de Petri na Modelagem de Workflows Científicos e seus Recursos*

Duílio Henrique Haroldo Elias, ***Kelly Rosa Braghetto***

Universidade de São Paulo / Instituto de Matemática e Estatística

duilio.elias@usp.br | kellyrb@ime.usp.br

Resumo

Nas últimas décadas os workflows científicos receberam grande atenção da comunidade acadêmica, devido à sua grande capacidade de manipulação, análise e simulação de experimentos que trabalham com grandes quantidades de dados. No entanto, as ferramentas existentes para tal fim são ainda escassas e muitos cientistas afirmam que esse seja o é um grande gargalo da produção científica. Neste trabalho, busca-se contribuir com este processo, criando modelos estocásticos capazes de descrever analiticamente alguns workflows bem estudado pela comunidade acadêmica. Com isso, pretende-se extrair índices capazes de predizer o impacto do uso de diferentes abordagens de controle de fluxo de dados no desempenho dos workflows.

Abstract

In the last decades, scientific workflows received great attention from the academic community, due to its large handling capability, analysis and simulation of experiments with large amounts of data. However, existing tools for this purpose are still scarce and many scientists claim that this is the major difficulty of the scientific production. In this work, we try to contribute to this process by creating stochastic models capable of describing analytically some workflows well studied by the academic community. With this, we intend to extract indexes able to predict the impact of using different approaches to control data flow in the performance of workflows.

*Este trabalho foi financiado por uma bolsa de iniciação científica RUSP (processo número: ?????).

Introdução

Um workflow científico(WfC) pode ser compreendido como um conjunto de tarefas computacionais que constituem um experimento científico [**?**], ou seja, os passos a serem executados pelo experimento. Os WfCs são conhecidos pelo seu poder de manipulação e execução de experimentos científicos automatizáveis que trabalham com grandes quantidades de dados o que os caracterizam como uma estruturas baseadas no fluxo de dados, ou seja, baseadas nas conexões entre as atividades e as dependência entre elas. Conceitualmente, um WfC pode ser visto como um grafo dirigido onde os nós são aplicações componentes e as arestas representam os fluxos de dados. Esse grafo pode ser acíclico, indicando que o WfC não contem controle de fluxo do tipo laço, onde um mesmo trecho do WfC poderia ser repetido até que alguma condição fosse satisfeita. Por *aplicações componentes*, entende-se como um programa de computador que implemente uma tarefa de um experimento científico computacional[**?**]. Um experimento pode ser composto por diversos WfC, que seguem um ciclo de vida.

As aplicações de WfCs são bastante extensas e a maioria delas visam automatizar processos computacionais complexos, assim como aumentar o desempenho dos mesmos.

As principais etapas do ciclo de vida de um WfC são a modelagem, execução e análise e são normalmente gerenciadas por sistemas complexos chamados de Sistemas Gerenciadores de Workflows Científicos(SGWfC). A especificação de um WfC pode ser através de uma *linguagem de workflow* ou por meio de uma *interface gráfica* que permita aos cientistas sem muito conhecimento avançados de computação especificar o workflow de forma mais fácil. Hoje existem diversos SGWfC e cada um deles apresentam particularidades para representação dos WfC que visam atender necessidades especificas de cada área. Os SGWfC mais utilizados, segundo o número compartilhamento no myexperiment[1], repositório de WfC, são:

- **Taverna**[2]: Tem como foco aparar cientistas da área de Ciências Biológicas é uma ferramenta multiplataforma e possue uma boa documentação. O Taverna possue uma interface gráfica que permite ao pesquisador construir WfC de forma muito parecida com fluxogramas e possuem suporte a *Web Services*.
- **Pegasus**[3]: O Pegasus é uma que permite aparar diversas áreas, permite a construção de WfC em sistemas distribuídos, também é uma solução multiplataforma, porém não possue uma interface gráfica intuitiva como o Taverna e os WfC são descritos em formato XML, na linguagem DAX que descreve as dependências entre atividades e os recursos computacionais, onde cada aplicação componente será executada.
- **Kepler** [4]: Assim como o Taverna o Kepler é uma ferramenta voltada para área de Bioinformática e os WfCs pode ser descritos com o auxílio de uma interface gráfica e também é uma ferramenta multiplataforma.

Estes são os principais SGWfC e estão descritos bem sucintamente, porém cada um tem suas particularidades que visam atender a necessidades especificas de cada área e cada cientista precisa escolher o que mais atende as suas necessidades.

Para este trabalho escolhemos o Pegasus, por ter uma boa documentação e permitir modelar sistemas com recursos computacionais distribuídos e também por possuir toda descrição em modo textual(DAX), que poderá no futuro criar um algoritmo para modelagem analítica de forma a automatizar criação de modelos analíticos que será feita nas próximas seções.

Objetivos

Detalhar um pouquinho as seguintes ideias:

- O objetivo do trabalhar é definir um método de conversão de modelos de workflows baseados em fluxos de dados em modelos em Rede de Petri Estocásticas
- Em modelos em Rede de Petri Estocásticas, pode-se também incorporar informações sobre o tempo de execução das atividades e também sobre os recursos computacionais disponíveis para a execução do workflow; por essa razão, é possível extrair deles predições sobre o desempenho do workflow.

[1] http://www.myexperiment.org/
[2] http://www.taverna.org.uk/
[3] http://pegasus.isi.edu/
[4] https://kepler-project.org/

Como a parte de comparar os resultados obtidos com as RdPs com os obtidos por meio do simulador não foi finalizada a tempo, não vamos mencioná-la aqui na seção de objetivos. Mas podemos falar dela na seção de trabalhos futuros.

O objetivo deste trabalho é definir um método de conversão de modelos de workflows baseados em fluxos de dados(WfC) em modelos em Rede de Petri Estocásticas. Para isso foram modelados um grupo de WfC, o qual se conhece bem o comportamento, utilizando para isso o arquivo DAX, neste arquivo é possível obter o tempo de execução de cada atividade, o tamanho de cada arquivo compartilhado e a precedência entre as atividades. Com essas informações foi possível construir as redes de Petri correspondentes. Em modelos em Rede de Petri Estocásticas, pode-se também incorporar informações sobre o taxa de execução das atividades e sobre os recursos computacionais disponíveis para a execução do WfC; por essa razão, é possível extrair deles predições sobre o desempenho do WfC.

Materiais e Métodos

Aqui a ideia é explicar os conceitos importantes relacionados ao seu trabalho e descrever de forma sucinta o que você fez:

- Explicar sucintamente os componentes de uma RdP
- Falar sobre as construções de fluxo de dados básicas (pipeline, distribuição, agregação)
- Para cada uma das construções acima, mostrar como ela é mapeada na RdP; também explicar que, no modelo em RdP, uma atividade é modelada como uma transição
- Dizer que, como estudo de caso, foi usado um conjunto de workflows científicos sintéticos (extraídos de workflows reais) que ilustram o uso de diferentes construções de fluxos de dados e que com frequência são empregados como base de comparação Citar a fonte desses workflows.
- Dizer que os workflows estavam descritos na linguagem DAX; seu modelos já trazem informações sobre o tempo médio de execução de cada atividade e também sobre o tamanho dos dados de entrada. Essas informações são importantes na construção do modelo estocástico em RdP
- Explicar como informações sobre os recursos computacionais disponíveis podem ser incluídas no modelo em RdP
- Falar que foi usada a ferramenta PIPE para a análise numérica do modelo.
- Dizer que a análise numérica do modelo nos dá de imediato dois índices de desempenho interessantes: a porcentagem de utilização dos recursos e o rendimento (throughput) das atividades e do workflow.

Resultados

Aqui não temos muita coisa para explicar. Temos que tentar incluir um ou dois daqueles workflows que você fez em Redes de Petri e depois dizer que os demais modelos estão no site (e colocar o link para aquela página que eu criei: `http://www.ime.usp.br/~kellyrb/ic/#duilio`).

Conclusões

Aqui é preciso dizer coisas como:

- No trabalho, foram modelados em RdP um conjunto de workflows científicos sintéticos (extraídos de workflows reais) que ilustram o uso de diferentes construções de fluxos de dados
- Concluiu-se que, por meio das Redes de Petri, é possível representar todas as construções básicas de fluxos de dados
- Recursos computacionais simples (como máquinas, VMs, processadores) são facilmente expressos nos modelos em RdP

- A solução numérica dos modelos em RdP traz uma predição do desempenho do workflow (utilização dos recursos e o rendimento (throughput) das atividades e do workflow)
- Dizer que, como trabalho futuro, intenciona-se comparar os resultados obtidos a partir das redes de petri com resultados obtidos por meio da simulação do workflow em ferramentas como o WorkflowSim.