Qualidade da Informação em Reformulação de Consultas em um PDMS: Uma Perspectiva

Bruno Felipe de França Souza[1], Ana Carolina Salgado[1], Maria da Conceição Moraes Batista[2]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Av. Professor Luís Freire, s/n, Cidade Universitária – 50740-540 – Recife – PE – Brasil

[2]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manoel de Medeiros, s/n, Dois Irmãos - CEP: 52171-900 – Recife – PE – Brasil

bffs@cin.ufpe.br, acs@cin.ufpe.br, ceca@deinfo@ufrpe.br

Abstract. *Information Quality (IQ) has been emerging as a key issue on information systems. It is possible to associate IQ aspects to dynamic and highly distributed systems. In a PDMS (Peer Data Management Systems), the use of IQ criteria may be a promising method to help query reformulation in a PDMS and to enrich the answers given to the user.*

Resumo. *Qualidade da Informação (QI) vem se tornando um aspecto crítico na área de sistemas de informações. Atrelado à sistemas dinâmicos e altamente distribuídos como um PDMS (Peer Data Management Systems), pode se mostrar um meio bastante promissor no que diz respeito à reformulação de consultas em um PDMS, enriquecendo assim a resposta dada para o usuário que submeteu um consulta em algum peer.*

1. Introdução

O processamento de dados distribuídos figura atualmente como uma das formas de disseminação e recuperação de dados mais usadas, principalmente no meio empresarial. Visando oferecer informações obtidas semanticamente, consultas mais fáceis, extração de dados de diversas fontes heterogêneas e um nível de abstração alto, muitas pesquisas estão voltadas aos sistemas considerados a evolução dos sistemas de integração de dados [Heese et al. 2005]: os sistemas chamados PDMS (*Peer Data Management Systems*).

PDMS são sistemas de gerenciamento de dados em ambientes *peer-to-peer* altamente voláteis, heterogêneos e distribuídos [Souza 2007]. Um PDMS (ver Figura 1) consiste em um conjunto de *peers* (físicos), cada *peer* tem o seu próprio esquema associado o qual representa seu domínio de atuação cujos dados são compartilhados através de mapeamentos entre peers. Um PDMS possui as seguintes características: compartilhamento descentralizado de dados; escalabilidade, processamento e armazenamento de dados feito a partir de *peers* autônomos, que também armazenam os mapeamentos semânticos dos dados [Neves 2008].

A Figura 1 mostra que fragmentos de cada *peer XML Schema* [Peterson et al. 2006] é exibido em forma de árvore com seus rótulos. As setas mostram que existem mapeamentos entre os esquemas dos *peers* [Tatarinov e Havely 2004].

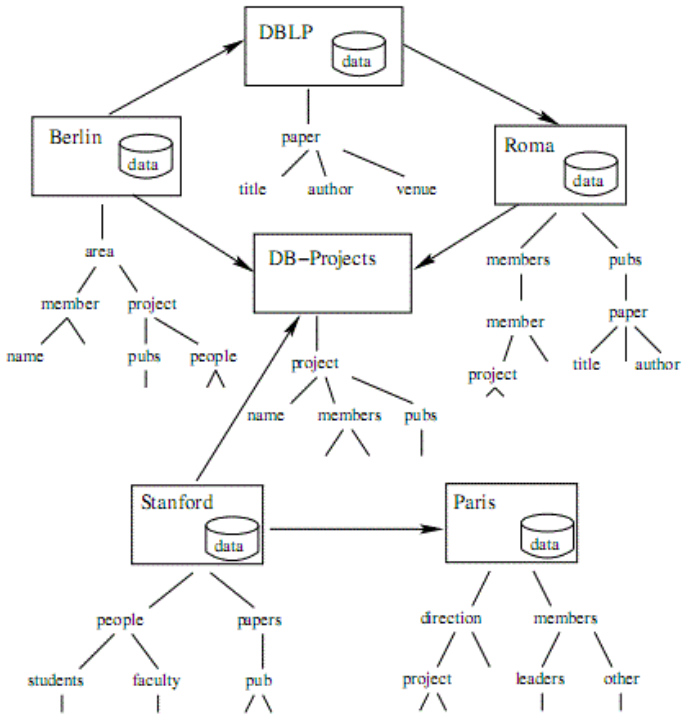

Figura 1 - Um PDMS para um banco de dados de pesquisa [Tatarinov e Havely 2004].

Qualidade da Informação (QI) é comumente definida como um conjunto de critérios para indicar o grau de qualidade geral de uma informação obtida por um sistema [Batista 2008]. A partir de artigos sobre qualidade da informação e qualidade dos dados (*data quality*) Naumann [Naumann 2001] chegou a uma verdade indiscutível afirmando que qualidade da informação é o principal fator discriminador de fonte de dados na *Web* e as métricas de qualidade da informação devem ser levadas em consideração para melhorar os resultados de consultas integradas. A partir desta definição, critérios de QI podem ser uma ferramenta bastante útil no enriquecimento de consultas através de sua reformulação, contribuindo assim para a melhora de problemas emergentes tais como: falta de disponibilidade dos *peers*; resultado de consultas incompletos.; tempo de resposta dos *peers* muito alto e inconsistência de conceito entre os *peers*.

Este artigo está organizado da seguinte forma: A Seção 2 discorre sobre critérios de QI. A Seção 3 descreve como é feito a reformulação de consultas e um PDMS. A seção 4 mostra uma perspectiva em relação à reformulação de consultas com critérios de QI. Na Seção 5, tem-se a conclusão do presente trabalho com algumas perspectivas.

2. Critérios de Qualidade da Informação

Alguns trabalhos de pesquisa consideram a qualidade da informação (QI) como sendo um dos aspectos mais importantes para integração de dados na *Web* [Batista 2003]. Informação de baixa qualidade é um dos problemas que mais perseguem os usuários de informações distribuídas por fontes de dados autônomas. Este cenário se torna mais forte para a variedade de tipos de usuários de informações da WWW [Nauman 2000].

Segundo Naumann [Naumann 2000], as métricas da qualidade da informação é a integração de aspectos da QI no processo de planejamento e otimização de consultas enviadas a um banco de dados ou sistema de informação. Aspectos de QI incluem um conjunto de critérios, métodos de avaliação e uma medição do grau da QI. Quando fontes de informações armazenam dados e informações sobre as mesmas entidades, aspectos de QI constituem a principal diferença entre as fontes de informações.

A qualidade da informação depende de três fatores maiores: a percepção do usuário, a informação em si e o acesso à informação [Naumann 2000] Os três fatores são classificados como o sujeito, o objeto e o predicado de uma consulta, e servem como um recurso para os metadados ou escores de QI. Um escore de QI é um valor associado a um determinado critério de QI.

Abaixo uma tabela contendo um conjunto de critérios de QI, estes voltados para sistemas de integração de dados.

Classe	Critério de QI	Método de Avaliação
Critérios de Sujeito	Possibilidade	Experiência do Usuário
	Representação Concisa	Amostragem do Usuário
	Interpretabilidade	Amostragem do Usuário
	Relevância	Avaliação Contínua do Usuário
	Reputação	Experiência do Usuário
	Compreensibilidade	Amostragem do Usuário
	Valor Agregado	Avaliação Contínua do Usuário
Critérios de Objeto	Completude	Parsing e Amostragem
	Suporte ao Usuário	Parsing e Contratação
	Documentação	Parsing
	Objetividade	Entradas de Usuário
	Preço	Contratação
	Confiabilidade	Avaliação Contínua
	Segurança	Parsing
	Atualidade	Parsing
	Rastreabilidade	Entradas de Usuário Especialista
Critérios de Processo	Precisão	Amostragem, técnicas de limpeza de dados
	Volume de Dados	Avaliação Contínua
	Disponibilidade	Avaliação Contínua
	Representação Consistente	Parsing
	Latência	Avaliação Contínua
	Tempo de Resposta	Avaliação Contínua

Tabela 1 - Classificação de critérios de QI [adaptada de Naumann 2000].

Dentro de cada classe é especificado o método de avaliação que deve ser aplicado para obtenção dos escores de cada um deles. Critérios subjetivos devem ser fixados pelo usuário por meio de métodos de experiência, amostragem e avaliação contínua.

Critérios objetivos podem ser avaliados automaticamente e apenas ocasionalmente entradas de usuários são necessárias, pode ser também avaliado de forma contínua por meio de *completude*.

Critérios de processos podem ser determinados através do processo de consultas, e assim variando de consulta para consulta e são representativos, porém temporários. Um exemplo de critério de processo seria *tempo de resposta*.

3. Reformulação de Consultas em PDMS

Um PDMS consiste em uma rede de nós denominados *peers*. *Peers* podem desempenhar um dos seguintes papéis: servidores de dados, mediadores para tradução entre esquemas de outros *peers* e pontos para execução de consultas [Tatarinov e Havely 2004].

O relacionamento entre os *peers* em um PDMS é dado através de mapeamentos, semânticos (ver [Sung et al. 2005]) entre os esquemas dos pares de *peers*. A Figura 2 mostra como consultas são disseminadas e traduzidas em um PDMS. Quando um usuário de **UB** (Universidade de Portugal) realiza uma consulta a primeira fonte a ser examinada em busca de dados é a própria **UB** (Universidade do Brasil). Consultas são processadas em **UB** e só então reformuladas e passadas para outros pontos vizinhos

através da rede de mapeamentos semânticos. Por exemplo, assumindo a existência de um mapeamento $\mathbf{Map_{UB\_UP}}$, entre **UB** e **UP**, a consulta $\mathbf{Q_{UB}}$ será reformulada para $\mathbf{Q_{UP}}$, de acordo com o esquema de **UP**. $\mathbf{Q_{UP}}$ será processada no esquema **UP**. Caso existam mapeamentos a consulta poderá ser reformulada para vizinhos adicionais a **UP**. Ao final, os resultados das consulta serão enviados ao ponto **UB** (inicial) e integrados depois das execuções nos pontos alcançados. Consequentemente o usuário receberá resultados não somente de **UB**, mas de todos os pontos que contribuíram com a resposta [Souza 2007].

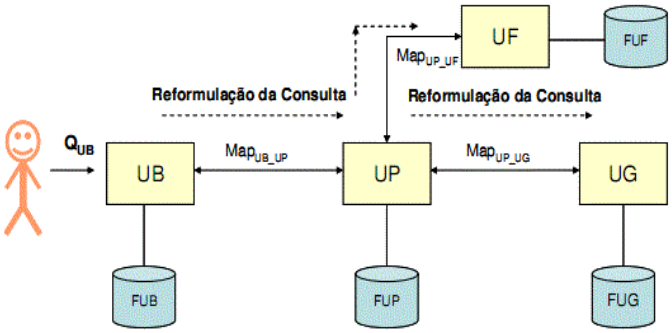

Figura 2 - Reformulação de consultas em um PDMS genérico, onde FUB - Fonte de Dados Universidade do Brasil [Souza 2007].

Pelo exemplo citado anteriormente, podemos concluir que a reformulação de consultas é um processo no qual uma consulta num esquema de fonte **A** é traduzida para o esquema da fonte de dados **B** de uma maneira que seja compreendida por **B**.

O processo de reformulação de uma consulta pode ser dividido em duas etapas: a reescrita da consulta que gera uma expressão de consulta (Q') e a resolução da consulta cujo resultado é o conjunto de todas as respostas possíveis para aquela expressão de consulta [Havely 2000].

4. Reformulação de Consultas com Critérios de Qualidade da Informação

Primeiramente, para atingirmos o objetivo desta seção, nós iremos fazer uma análise minuciosa sobre os critérios de QI e seus impactos na reformulação de consultas. Alguns desses critérios podem ser extraídos da compilação feita por Naumann que foi apresentada na Tabela 1 da Seção 3 bem como em pesquisas no estado da arte com relação à QI.

Ainda que sem uma investigação profunda podemos citar alguns critérios listados por Naumann na Tabela 1 e Wang [Wang 1996], como sendo provavelmente relevantes em reformulação de consultas em um PDMS, devido a estes critérios serem de tamanha importância para uma melhora significativa na qualidade de diversos tipos de sistemas computadorizados. A Tabela 2 mostra a relação entre os elementos presentes em um PDMS com seus critérios de QI associados.

Elementos de um PDMS	Critérios de QI
Fonte de dados (*peers*)	Reputação, Fácil Acesso, Tempo de Resposta, Disponibilidade
Esquema (compartilhado por cada *peer*)	Integridade do Esquema, Minimalidade, Representação Consistente, Completude
Mapeamentos (entre *peers*)	Completude, Precisão, Atualidade
Dados	Confiabilidade, Objetividade, Precisão, Integridade

Tabela 2 – Elementos em um PDMS e critérios de QI.

Usando esses escores de QI relacionados com os elementos de um PDMS cada *peer* irá encontrar pontos relevantes à consulta e processá-la eficientemente.

Dada uma fonte de dados **FD**, um conjunto de critérios de **QI** entre os *peers* e uma consulta de usuário **Q**, busca-se encontrar uma consulta **Q'** de **Q** usando **QI** como um conjunto de critérios de qualidade da informação de modo que **Q'** retorne respostas mais significativas ao usuário da consulta **Q**.

6. Conclusão

O presente artigo pretendeu levantar a discussão acerca do uso da QI em sistemas de acesso à informações, mais especificamente sistemas PDMS. Como mencionado neste trabalho, o processos de reformulação de consultas pode ser auxiliado através do uso e avaliação de critérios de QI. Nossos trabalhos futuros consistem em: investigar minuciosamente quais critérios de QI podem efetivamente ser usados para auxiliar na reformulação de consultas distribuídas; Especificar formalmente os critérios selecionados; Implementar a avaliação destes critérios em consultas de um ambiente PDMS; Avaliar os resultados do uso de QI em consultas em ambiente PDMS.

7. Referências

[Tatarinov, I. 2004], Havely, A.Y. "Efficient Query Reformulation in Peer Data Management Systems" In SIGMOD 2004 – Paris – France.

[Batista, M. C. M. 2003] "Otimização de Acesso em Um Sistema de Integração de Dados através do Uso de Caching e Materialização de Dados", Dissertação de Mestrado – UFPE.

[Batista, M. C. M. 2008] "Schema Quality Analysis in a Data Integration System". Tese de Doutorado, Centro de Informática – UFPE.

[Naumann, F. 2000] and Rolker, C. "Assessment Methods for Information Quality Criteria". In Proceedings of the Conference on International Quality (IQ00) Boston, 2000.

[Naumann, F. 2001] "From Databases to Information Systems – Information Quality Makes the Difference". In 6th International Conference on Information Quality (IQ01) Boston, 2001.

[Neves, T. A. 2008] "Desenvolvimento do Módulo de Reformulação de Consultas no Sistema SPEED". Federal University of Pernambuco (UFPE/CIn). Undergraduate Conclusion Monograph. Recife, PE, Brazil.

[Halevy, A. Y. 2000]. "Theory of Answering Queries Using View". ACM Special Interest Group on Management of Data Record 29(4), 40--47.

[Heese, R. 2005], Herschel S., Naumann F., and Roth A. (2005) "Self-extending peer data management". In G. Vossen, F. Leymann, P. C. Lockemann, and W. Stucky, editors, Proceedings of the German Conference on Datenbanksysteme in Business, Technologie und Web, volume 65 of LNI. GI, March 2005.

[Peterson et al. 2006] Peterson, D., Biron, P. V., Malhotra, A. and Sperberg- McQueen., C. M. XML Schema 1.1 Part 2: Data Types – W3C Working Draft, http://www.w3.org/TR/xmlschema11-2/, 2006. Acessado em 19 de março de 2011.

[Souza, D.Y. 2007]"Reformulação de Consultas Baseadas em Semântica para PDMS, Exame de Qualificação e Proposta de Tese – UFPE.

[Sung, L. G. A. 2005], Ahmed, N., Blanco, R., Li, H, Soliman, M. A., and Hadaller, D. "A Survey of Data Management in Peer-to-Peer Systems". School of Computer Science, University of Waterloo.

[Wang, R.Y. 1996] and Strong, D. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management of Information Systems, 12, 4: pp.5-34, 1996.