Avaliando a Relevância de Novas Fontes de Dados em Sistemas de Integração de Dados *Pay-as-you-go*

Aluno: Hélio Rodrigues de Oliveira
hro@cin.ufpe.br
Orientadora: Bernadette Farias Lóscio
bfl@cin.ufpe.br

Nível: Mestrado
Ano de Ingresso: 2011
Data da defesa da proposta: março de 2012
Previsão de Conclusão: março de 2013
Etapas já concluídas: Revisão bibliográfica, definição do problema e escrita da proposta
Etapas futuras: Defesa da proposta, especificação e implementação da solução, escrita da dissertação, defesa e publicação de resultados

Programa de Pós Graduação em Ciência da Computação
Universidade Federal de Pernambuco (UFPE)
Recife – Pernambuco – Brasil

Resumo

Diferentemente dos sistemas de integração convencionais, os sistemas de integração de dados pay-as-you-go são caracterizados pelo baixo custo de inicialização do sistema, porém com baixa qualidade no resultado das consultas. Por meio da integração sob demanda, utilizando o feedback do usuário, os mapeamentos entre as fontes de dados e os esquemas de mediação são refinados até que os resultados alcancem os requisitos do usuário. Nesse contexto, considerando que novas fontes de dados possam ser adicionadas com mais facilidade, o sistema torna-se mais flexível e dinâmico. Porém, devido à heterogeneidade das fontes de dados, muitas delas podem conter informações irrelevantes para o sistema de integração. Uma das formas para melhorar a qualidade dos resultados das consultas é minimizar a quantidade de fontes "sujas" adicionadas ao ambiente. Assim, propomos uma abordagem para análise da relevância de novas fontes de dados em ambientes de integração pay-as-you-go. Utilizando o feedback do usuário relacionado a um conjunto de consultas, poderemos avaliar se uma fonte candidata a entrar no sistema é ou não relevante para este conjunto de consultas.

Palavras-Chave

Pay-as-you-go, feedback do usuário, qualidade da informação.

1. Introdução e Motivação

Integração de dados é uma importante área de pesquisa que evoluiu bastante com o surgimento da Web e a crescente facilidade de acesso a grandes volumes dados. De uma maneira geral, o objetivo da integração de dados é permitir o acesso, através de uma visão unificada, a um conjunto de fontes de dados autônomas e heterogêneas [Lóscio 2003]. Os sistemas de integração de dados convencionais geralmente possuem um esquema de mediação, que representa a visão unificada dos dados, e um conjunto de mapeamentos, que relacionam o esquema de mediação com os esquemas das fontes de dados. Como passo de inicialização no processo de integração, temos a identificação das fontes de dados, a geração do esquema de mediação e dos mapeamentos entre esquemas. Esta inicialização tem sido apontada como um dos principais gargalos dos tradicionais sistemas de integração de dados [McCann et al. 2005], principalmente, por se tratar de um processo manual ou semiautomático.

Como uma solução para este problema surgiram os sistemas de integração de dados *pay-as-you-go* [Vaz Salles et al. 2007; Das Sarma et al. 2008; Wang et al. 2009], os quais tem o objetivo de prover os benefícios da integração de dados clássica, porém fazem uso de uma abordagem de integração de dados sob demanda, onde os mapeamentos entre os esquemas são gerados de forma automática e incremental, podendo, além disso, sofrer refinamentos em tempo de execução. Dessa forma, é possível reduzir o custo de inicialização do sistema. Entretanto, vale ressaltar que, neste caso, os resultados das consultas podem ser menos precisos.

Neste contexto de ambientes dinâmicos e flexíveis, é importante observar que a adição de fontes, facilitada pelo baixo custo de manutenção, aumenta a chance de inclusão de dados pouco relevantes ao sistema, podendo contribuir para a geração de resultados ineficientes para algumas consultas. Minimizar a quantidade destas fontes "sujas" carregadas pelo ambiente é uma das maneiras possíveis para melhorar a qualidade dos resultados oferecidos pelo sistema de integração de dados *pay-as-you-go*.

Neste trabalho, propomos uma abordagem para analisar a influência da adição de novas fontes de dados a um sistema de integração de dados *pay-as-you-go*. Esta análise propõe identificar se a adição da nova fonte irá contribuir para a geração de melhores resultados de consultas, em comparação com os resultados obtidos previamente, ou seja, antes da inserção desta fonte. Considerando um conjunto de consultas e o *feedback* do usuário relacionado a este conjunto, seremos capazes de analisar o quão relevante pode ser a nova fonte para o sistema como um todo.

2. Fundamentação teórica

Os sistemas de integração convencionais tem uma visão "primeiro o esquema, depois os dados", ou seja, investem, inicialmente, uma grande quantidade de recursos para que o esquema de mediação e os mapeamentos sejam bem definidos, e somente depois oferecem o acesso integrado aos dados distribuídos. Dessa forma, o resultado das consultas sobre os dados integrados são mais precisos, exigindo, porém um alto custo de inicialização.

Por outro lado, os sistemas de integração de dados *pay-as-you-go* proveem baixo custo de inicialização por meio da geração incremental e automática de mapeamentos e esquemas de mediação. Porém, como consequência desta abordagem, os resultados recuperados pelas consultas podem ser menos precisos. Como forma de incrementar a precisão, utilizamos o *feedback* do usuário. O *feedback* pode ser visto como um conjunto de anotações que o usuário provê sobre artefatos do sistema de integração, sejam consultas, mapeamentos, esquema de mediação, entre outros, com o objetivo de melhorar

a qualidade dos serviços disponibilizados [Belhajjame et al. 2011]. Estas anotações exprimem de forma mais abstrata quais são os requisitos do usuário sobre o artefato anotado. A preocupação em utilizar o *feedback* em ambientes *pay-as-you-go* tem se tornado evidente, como mostram algumas pesquisas na área [Chai et al. 2009; Cao et al. 2010].

Em particular, os sistemas de integração de dados *pay-as-you-go* utilizam o *feedback* do usuário sobre cada consulta realizada no sistema [Jeffery et al. 2008]. As anotações obtidas são, em geral, utilizadas para permitir o refinamento incremental dos mapeamentos entre esquemas, através da reexecução das consultas, considerando as anotações realizadas previamente sobre elas. O processo de refinamento é realizado até que os mapeamentos sejam capazes de atingir os requisitos do usuário.

3. Caracterização da Contribuição

Neste trabalho propomos uma abordagem para análise da relevância de novas fontes de dados em ambientes de integração de dados que adotam uma abordagem *pay-as-you-go*. De maneira mais específica, estamos interessados em sistemas de integração de dados que levam em consideração o *feedback* fornecido pelo usuário sobre as consultas submetidas aos sistemas [Belhajjame et al. 2011].

Consideramos que o *feedback* está relacionado a um conjunto de consultas, onde cada uma das consultas está associada a um peso, que determina a sua relevância para o sistema de integração. Neste trabalho, o peso de cada consulta é dado pela sua frequência de execução, ou seja, quanto maior o número de execuções de uma consulta, mais relevante esta será para o sistema.

Em nossa abordagem, assumimos que um sistema de integração de dados *pay-as-you-go* pode prover um conjunto de informações necessárias, denominado de configuração, o qual será utilizado em nossa proposta. De maneira mais específica temos que, uma configuração T obtida a partir de um sistema de integração de dados *pay-as-you-go* pode ser definida pela quádrupla $\mathrm{T} = (\mathrm{S}, \mathrm{Q}, \mathrm{F}, \mathrm{W})$, onde $\mathrm{S} = \{\mathrm{s}_1, \mathrm{s}_2, \dots, \mathrm{s}_\mathrm{n}\}$ é um conjunto de fontes de dados, $\mathrm{Q} = \{\mathrm{q}_1, \mathrm{q}_2, \dots, \mathrm{q}_\mathrm{m}\}$ um conjunto de consultas, tal que q_i é uma consulta realizada sobre uma ou mais fontes de S, $\mathrm{F} = \{\mathrm{f}_1, \mathrm{f}_2, \dots, \mathrm{f}_\mathrm{m}\}$ um conjunto de *feedbacks*, tal que f_i é o *feedback* relativo à consulta q_i em Q e $\mathrm{W} = \{\mathrm{w}_1, \mathrm{w}_2, \dots, \mathrm{w}_\mathrm{n}\}$ um conjunto de pesos, tal que w_i é o peso relativo à consulta q_i. O problema da análise da relevância de uma nova fonte para um sistema de integração *pay-as-you-go* pode ser definido como se segue:

Dada uma configuração $T = (S, Q, F, W)$, obtida a partir de um sistema de integração de dados pay-as-you-go I, e uma nova fonte de dados s, devemos descobrir qual a relevância de s sobre I, com relação às consultas em Q. Consideramos que uma fonte de dados s é relevante, com respeito à Q, se os resultados obtidos sobre este conjunto de consultas forem mais precisos, de acordo com os requisitos do usuário, após a adição de s.

Uma das formas de medir a relevância das fontes com relação às consultas consiste em reutilizar as anotações obtidas pelo *feedback* do usuário. Estas anotações podem ser classificadas em um dos seguintes casos: i) resultados esperados pelo usuário (*true positives* ou TP); ii) resultados que não estão de acordo com os requisitos do usuário (*false positives* ou FP); e iii) resultados omissos (*false negatives* ou FN).

Estas anotações serão utilizadas em nosso trabalho para o cálculo dos valores de *precision* e *recall* obtidos do *feedback* para cada consulta Q. De maneira mais específica,

temos que *Precision* é a razão entre o número de *true positives* e a soma de *true positives* e *false positives*. Similarmente, *recall* é a razão do número de *true positives* e a soma de *true positives* e *false negatives*.

Para ilustrar nossa abordagem, considere o cenário descrito a seguir. Seja I um sistema de integração construído sobre o domínio de informações (geográficas, políticas e econômicas) de países. Construímos T a partir de I da seguinte forma: $S = \{s_1, s_2, s_3\}$, $Q = \{q_1, q_2, q_3, q_4\}$, $F = \{f_1, f_2, f_3, f_4\}$ e $W = \{w_1, w_2, w_3, w_4\}$. A fonte s_1 possui informações sobre países europeus, s_2 sobre países africanos e s_3 informações sobre países de uma maneira geral. As consultas q_1, q_2, q_3 e q_4 são consultas já executadas no sistema, com *feedbacks* f_1, f_2, f_3 e f_4, e pesos w_1, w_2, w_3 e w_4, respectivamente. Em nosso exemplo, utilizaremos a consulta q_1 para mostrar o processo de análise, no qual o procedimento realizado é análogo para todas as outras consultas.

Suponha que a consulta q_1 seleciona informações geográficas sobre os países. Porém, considere que o usuário deseja obter informações apenas dos países europeus e mediterrâneos. Ao executar a consulta q_1, foi retornado um conjunto de resultados, e estes foram anotados pelo usuário, como podemos ver na Tabela 1. Observe que são obtidas anotações sobre dados esperados, não esperados e resultados omissos.

Tabela 1. Feedback f_1 relativo a consulta q_1.

Nome	Capital	Área	População	Feedback
Turkey	Ankara	780580	62484478	TP
Greece	Athens	131940	10538594	TP
Spain	Madrid	504750	39181114	TP
Denmark	Copenhagen	43070	5249632	FP
Germany	Berlin	356910	83536115	FP
Egypt	Cairo	1001450	63575107	FP
Iran	Tehran	16480000	66094264	FP
Monaco	Monaco	-	-	FN
Italy	Rome	-	-	FN
Malta	Valletta	-	-	FN

Suponha agora a adição de uma nova fonte s_4, que define informações sobre países mediterrâneos. Queremos avaliar se s_4 é relevante para o sistema I, com relação à Q. O processo de análise consiste nos seguintes passos:

Inicialmente, calculam-se os valores de *precision* e *recall* relacionados ao *feedback* f_1 de q_1. Neste caso, obtemos $precision(f_1) = 0.43$ e $recall(f_1) = 0.5$. A seguir, reexecutamos a consulta q_1, incluindo também a fonte s_4, e obtemos um novo conjunto de resultados para q_1. A partir do *feedback* anterior, inferimos anotações sobre o novo conjunto de respostas. Os resultados da consulta e o novo *feedback* f_1' são mostrados na Tabela 2. O próximo passo do processo de análise consiste em calcular os novos valores de *precision* e *recall* de acordo com as anotações inferidas. Assim, temos como novos valores: $precision(f_1') = 0.83$ e $recall(f_1') = 0.71$. Como resultado para a consulta q_1, a nova fonte s_4 resultou em uma taxa de aumento de 93% no *precision* e de 42% no *recall*, ou seja, com respeito à consulta q_1, a adição de s_4 melhorou os resultados da consulta, de acordo com os requisitos do usuário.

É importante observar que os resultados obtidos ilustram a relevância de uma nova fonte para apenas uma consulta específica. Sendo assim, devemos repetir o processo para todas as consultas em Q, e assim calcular a relevância total da fonte para o sistema como

um todo. O cálculo da relevância total é dado pela média ponderada das relevâncias de cada consulta considerando os pesos em *W*.

Tabela 2. Feedback inferido f_1' utilizando as anotações da tabela 1.

Nome	Capital	Área	População	Feedback
Turkey	Ankara	780580	62484478	TP
Greece	Athens	131940	10538594	TP
Spain	Madrid	504750	39181114	TP
Monaco	Monaco	1.9	31719	TP
Gibraltar	Gibraltar	6.5	28765	TP
Egypt	Cairo	1001450	63575107	FP
Italy	Rome	-	-	FP
Malta	Valletta	-	-	FN

4. Trabalhos Relacionados

Alguns trabalhos tratam questões que serão abordadas em nossa proposta, como, por exemplo, [Wang et al. 2009] que propõe a utilização de dependências funcionais para a descoberta de fontes "sujas", incompletas ou mal interpretadas. Neste trabalho são definidas as (*p*FD) dependências funcionais probabilísticas, que permitem realizar a descoberta de dependências funcionais em sistemas de integração *pay-as-you-go*. Observando a qualidade das dependências funcionais durante a adição de novas fontes de dados, é possível decidir se a fonte é suficientemente boa para o sistema ou não. Nosso trabalho difere deste, pois, utilizamos informações obtidas do *feedback* do usuário (*precision* e *recall*) para realizar uma análise da relevância da adição de uma nova fonte, e decidir se esta fonte é ou não relevante para o sistema.

Com respeito ao *feedback* do usuário, [Talukdar et al. 2010] propõe uma abordagem para adição de novas fontes utilizando buscas por palavras-chaves. Inicialmente, é gerado um grafo de busca a partir das fontes de dados e de seus relacionamentos. A busca por palavra-chave é realizada sobre o grafo, e os resultados são retornados em uma visão *top-k* contendo as respostas mais relevantes para o usuário. A manutenção do grafo é feita incrementalmente através do *feedback*, e quando novas fontes são descobertas, o grafo é realinhado. Porém a abordagem citada utiliza um modelo de integração voltada à necessidade de informação baseado no sistema Q [Talukdar et al. 2008]. Em nosso trabalho, a abordagem proposta pode ser aplicada em ambientes de integração de dados dinâmicos e flexíveis, desde que o mesmo permita obter as configurações necessárias, definidas na seção 3.

5. Avaliação dos Resultados e Estado Atual do Trabalho

A validação da abordagem proposta neste trabalho inclui a implementação de um protótipo, juntamente com a realização de um conjunto de experimentos, a fim de avaliar principalmente se uma fonte, que não é relevante para o sistema de integração, pode degradar a qualidade dos resultados das consultas. Desta forma, pretendemos mostrar que ao realizar a análise desta fonte, evitamos perda na qualidade dos resultados das consultas caso a fonte fosse inserida no sistema. Além disso, devemos mostrar como a utilização da abordagem pode contribuir para a redução do custo da manutenção incremental dos mapeamentos, uma vez que contribui para redução da quantidade de *feedback* necessário para refiná-los.

Atualmente, o trabalho encontra-se em fase de fundamentação teórica e definição dos cálculos para medição da relevância de novas fontes. Por outro lado, estamos

trabalhando em um estudo de caso utilizando dados sintéticos para obtenção de resultados preliminares. Posteriormente, com o intuito de validar nossa abordagem, pretendemos aplicá-la no contexto de Linked Open Data.

Ao final deste trabalho, as principais contribuições esperadas são a redução do custo necessário para manutenção incremental em sistemas de integração de dados *pay-as-you-go*, bem como a proposta e implementação de um protótipo para validação da análise da influência das novas fontes de dados em sistemas de integração *pay-as-you-go*.

Referências

Belhajjame, K., Paton, N. W., Fernandes, A. A. A. Hedeler, C., Embury, S. M. (2011). "User Feedback as a First Class Citizen in Information Integration Systems". In: Proceedings of the 5th Conference on Innovative Data Systems Research, CIDR 2011, 175-183.

Cao, H., Qi, Y., Candan, K. S., Sapino, M. L. (2010). "Feedback-driven Result Ranking and Query Refinement for Exploring Semi-structured Data Collections". In: Proceedings of the 13th International Conference on Extending Database Technology, EDBT 2010, 3 – 14.

Chai, X., Vuong, B. Q., Doan, A., Naughton, J. F. (2009)."Incorporating User Feedback into Information Extraction and Integration Programs". In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, 87-100.

Das Sarma, A., Dong, X., Halevy, A. (2008). "Bootstrapping Pay-as-you-go Data Integration Systems". In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, 861–874.

Jeffery, S. R., Franklin, M. J., Halevy, A. Y. (2008). "Pay-as-you-go User Feedback for Dataspace Systems". In: Proceedings of the of ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, 847-860.

Lóscio, B. F. (2003). "Managing the Evolution of XML-based Mediation Queries". PhD thesis, Informatics Center, Federal University of Pernambuco, Recife, Brasil, 2003. .

McCann, R., AlShebli, B. K., Le, Q., Nguyen, H., Vu, L. Doan, A. (2005). "Mapping Maintenance for Data Integration Systems". In: Proceedings of Very large Data Bases Conference,VLDB 2005, 1018–1030.

Talukdar, P. P., Ives, Z. G., Pereira, F. (2010). "Automatically Incorporating New Sources in Keyword Search-based Data Integration". In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, 387-398.

Talukdar, P. P., Jacob, M., Mehmood, M. S., Crammer, K., Ives, Z. G., Pereira, F., Guha, S. (2008). "Learning to Create Data-integrating Queries". In: Proceedings of Very large Data Bases Conference,VLDB 2008, 785-796.

Vaz Salles, M.A., Dittrich, J. P., Karakashian, S. K., Girard, O. R., Blunschi, L. (2007). "iTrails: Pay-as-you-go Information Integration in Dataspaces". In: Proceedings of Very large Data Bases Conference, VLDB 2007, 663-674.

Wang, D. Z., Dong, X. L., Das Sarma, A., Franklin, M. J., Halevy, A. Y. (2009). "Functional Dependency Generation and Applications in Pay-As-You-Go Data Integration Systems". In: Proceedings of the 12th International Workshop on the Web and Databases, WebDB 2009.