Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas

Everaldo Costa Silva Neto[1], **Bernadette Farias Lóscio**[1], **Ana Carolina Salgado**[1]

[1]Centro de Informática - Universidade Federal de Pernambuco (UFPE)
Recife - Pernambuco - Brasil

{ecsn, bfl, acs}@cin.ufpe.br

Abstract. *Nowadays, a massive volume of data has been produced by a variety of data sources. The easy access to these data presents new opportunities. In this sense, choosing the most suitable data sources for a specific use has become a challenge. The literature contains many works that perform quality assessment in data sources as a mean of solving this issue. However, only few works take into account the dynamicity of sources. In this work, we address the problem of performing data quality assessment in dynamic data sources. Furthermore, we propose the establishment of a Quality Profile, which consists in a set of metadata that provides information about the quality of a data source. The experiments performed on real-world scenarios have demonstrated that our strategy produces satisfactory results.*

Resumo. *Atualmente, um massivo volume de dados tem sido produzido pelos mais variados tipos de fontes de dados. A facilidade de acesso a esses dados apresenta novas oportunidades, no entanto, escolher quais fontes de dados são mais adequadas para um determinado uso tornou-se um desafio. A literatura oferece diversos trabalhos que abordam a avaliação da qualidade em fontes de dados como meio para solucionar esse desafio, entretanto, poucos trabalhos consideram o aspecto dinâmico das fontes. Neste trabalho, abordamos o problema de avaliação da qualidade em fontes de dados dinâmicas. Além disso, propomos a criação de um Perfil de Qualidade, que consiste de um conjunto de metadados que oferece dados sobre a qualidade de uma fonte e que poderá ser utilizado para facilitar o processo de seleção de fontes de dados. Os experimentos realizados demonstraram que a estratégia de avaliação da qualidade proposta produz resultados satisfatórios.*

1. Introdução

Nos últimos tempos, um volume crescente de dados tem sido produzido pelos mais variados tipos de fontes de dados, incluindo desde sistemas transacionais até dispositivos móveis. Além disso, iniciativas como a publicação de Dados na *Web* [Lóscio et al. 2016] e o paradigma de WoT (*Web of Things*) [Duquennoy et al. 2009] tornam o acesso a esses dados cada vez mais fácil. Entretanto, se por um lado o grande volume e a facilidade de acesso aos dados apresenta novas oportunidades, por outro lado escolher as fontes de dados que são mais adequadas para um determinado uso tornou-se um desafio. Normalmente, um conjunto de características, que vão desde aspectos ligados à proveniência [Malaverri et al. 2014] e qualidade dos dados [Xian et al. 2009, Lóscio et al. 2012], até a qualidade do serviço que é provido pelas fontes de dados [Dustdar et al. 2012], são utilizados para a escolha das fontes de dados.

É importante destacar que a literatura oferece diversos trabalhos que abordam a avaliação da qualidade de fontes de dados como meio para solucionar o problema da seleção de fontes [Xian et al. 2009, Lóscio et al. 2012, Dong et al. 2013]. Em geral, esses trabalhos propõem o uso de critérios da Qualidade da Informação (QI) [Wang and Strong 1996], tais como Disponibilidade, Corretude e Completude, para a avaliação das fontes de dados. Uma característica comum desses trabalhos é que a avaliação da qualidade de uma fonte de dados é realizada de forma pontual, ou seja, considerando apenas um instante de tempo específico. Apesar de ser uma abordagem bastante utilizada, esse tipo de avaliação pode levar a valores de qualidade muito além ou aquém dos valores reais de qualidade da fonte. Isso acontece porque os valores de qualidade de uma fonte podem variar de acordo com mudanças na própria fonte de dados, bem como de acordo com mudanças no ambiente. Dessa forma, torna-se necessário realizar uma avaliação que leve em consideração não apenas o estado atual da fonte de dados, mas também valores que representem a qualidade da fonte em momentos anteriores.

Neste trabalho, abordamos o problema de avaliação da qualidade de fontes de dados dinâmicas, ou seja, fontes de dados cujo conteúdo pode sofrer modificações com alta frequência [Rekatsinas et al. 2014]. A fim de considerar o aspecto dinâmico das fontes no processo de avaliação da qualidade, propomos uma estratégia de avaliação contínua. A estratégia proposta considera um conjunto de critérios de QI e, para cada critério, prevê uma forma de avaliação com base no estado atual da fonte de dados e nos valores de qualidade obtidos anteriormente.

Além disso, propomos a criação de um Perfil de Qualidade para as fontes de dados. O Perfil de Qualidade consiste de um conjunto de metadados que poderão ser disponibilizados juntamente com a fonte de dados, e poderá ser utilizado para facilitar o processo de seleção de fontes de dados, por exemplo. Ademais, propomos que o Perfil de Qualidade seja atualizado periodicamente, segundo a frequência de atualização da fonte de dados. Dessa forma, esperamos que o Perfil de Qualidade de uma fonte possa refletir, de forma mais precisa, a qualidade da fonte.

A fim de avaliar a nossa proposta, realizamos experimentos com dados meteorológicos disponibilizados por instituições que monitoram as condições climáticas da cidade do Recife (ITEP[1] e APAC[2]). Esses dados são gerados por sensores, os quais podem ser considerados fontes de dados dinâmicas. Os resultados obtidos nos experimentos demonstraram que a estratégia de avaliação contínua da qualidade dos dados produz melhores resultados do que a avaliação pontual da qualidade de dados.

O restante do trabalho está organizado da seguinte forma. A Seção 2 apresenta um exemplo motivacional e a definição do problema. A Seção 3 destaca algumas definições preliminares que serão utilizadas para especificar o processo de geração do Perfil de Qualidade, que será apresentado na Seção 4. A Seção 5 apresenta alguns resultados obtidos por meio dos experimentos realizados. A Seção 6 destaca os trabalhos relacionados e, por fim, a Seção 7 apresenta as conclusões e indicações de trabalhos futuros.

2. Definição do Problema

Nesta seção, apresentamos um exemplo que ilustra o problema de avaliação da qualidade de fontes de dados dinâmicas, ou seja, fontes que sofrem atualizações com alta frequência.

[1]http://www.itep.br/

[2]http://www.apac.pe.gov.br/meteorologia

É importante destacar que, neste trabalho, apenas as inclusões de novos dados são tratadas como atualizações da fonte de dados.

Considere um sensor (S) que coleta dados climáticos de uma região da cidade de Recife, com uma frequência predeterminada, armazenando-os em lote. A cada 24 horas, os dados coletados são armazenados em um banco de dados e, posteriormente, disponibilizados em formato aberto para uso pelo público em geral. Suponha que a Secretaria de Meio Ambiente deseja desenvolver uma aplicação para ofecer análises sobre dados climáticos da cidade. Para isso, devem ser selecionadas as fontes de dados mais adequadas. Dessa forma, o responsável pelo desenvolvimento da aplicação faz uma avaliação da qualidade das fontes de dados disponíveis, incluindo a fonte que disponibiliza os dados coletados a partir do sensor descrito acima. De maneira específica, o desenvolvedor deseja avaliar: (i) a fonte oferece dados corretos? e (ii) a fonte é completa, ou seja, a fonte oferece todas as instâncias de dados esperadas para um intervalo de tempo específico?

Uma maneira de descrever com precisão a qualidade da fonte é, a cada nova inserção de dados, realizar uma nova avaliação incluindo as novas entradas de dados. Para o nosso exemplo, como o banco de dados sofre atualizações a cada 24 horas, então uma nova avaliação teria que ser feita a cada 24 horas. Entretanto, o custo computacional associado a tal ação pode ser alto. À medida que novos dados são coletados pelo sensor e adicionados ao banco de dados, o volume de dados tende a crescer muito, de tal forma que a avaliação de qualidade do conjunto de dados completo pode ser custosa. O processo pode ficar ainda mais caro quando a avaliação contemplar múltiplos critérios de qualidade.

A fim de tratar o problema descrito anteriormente, algumas soluções encontradas na literatura consideram apenas um intervalo de tempo específico ou um subconjunto dos dados no momento da avaliação da qualidade de uma fonte de dados. A abordagem proposta por Xian et al. [Xian et al. 2009], por exemplo, seleciona randomicamente uma amostra dos dados contidos na fonte para realizar a avaliação da qualidade. Porém, optar por uma estratégia desse tipo, para o cenário que descrevemos, pode não refletir com precisão a qualidade real da fonte de dados. É importante lembrar que problemas durante a coleta dos dados, como perda de dados ou entrada de dados incorretos, podem afetar a qualidade dos dados. Dessa forma, dependendo da amostra escolhida, a precisão da avaliação da qualidade dos dados poderá ser prejudicada.

Sendo assim, surge a seguinte questão de pesquisa: *Como avaliar a qualidade de uma fonte de dados, que sofre atualizações de inserção de dados com uma frequência elevada, considerando que: (i) a qualidade dos dados poderá sofrer atualizações ao longo do tempo e (ii) um grande volume de dados será gerado a partir das frequentes inserções de dados?*

3. Perfil de Qualidade

Na literatura, o conceito de QI é multidimensional, ou seja, é baseado em um conjunto de critérios (ou dimensões), onde o papel de cada um é avaliar e medir um aspecto específico da qualidade. Cada critério pode estar associado a uma ou mais métricas. As métricas são heurísticas desenvolvidas para se adequar a uma situação de avaliação específica [Pipino et al. 2002]. Com relação às métricas, elas podem ser objetivas (quando são mensuradas de forma automática) ou subjetivas (quando necessita da intervenção humana para que possa ser mensurado um valor para um determinado critério).

Neste artigo, usamos critérios de QI para a definição de um Perfil de Qualidade para fontes de dados. O Perfil de Qualidade de uma fonte de dados f consiste de um conjunto de metadados que descrevem a qualidade da fonte em termos de um conjunto de critérios de QI, $Q = \{Q_1, ..., Q_m\}$, tal que Q denota o conjunto de critérios usados na avaliação de qualidade da fonte f.

O Perfil de Qualidade de uma fonte pode ser usado, por exemplo, para auxiliar um usuário a avaliar se a fonte de dados se adequa para seu uso, ou até mesmo auxiliar no processo de selecionar fontes para uma atividade específica. Além disso, os benefícios oferecidos pelo Perfil de Qualidade não se limitam apenas aos consumidores dos dados, o provedor dos dados também pode ser beneficiado, uma vez que é possível acompanhar a evolução da qualidade por meio da análise do perfil.

3.1. Definições Preliminares

Nesta seção, fornecemos algumas definições preliminares para auxiliar no entendimento do processo de geração do Perfil de Qualidade.

Definição 3.1. Conjunto de Valores dos Critérios de Qualidade: O conjunto CQ consiste de um conjunto de pares, $\{(Q_1, v_1), ..., (Q_m, v_m)\}$, onde para cada par (Q_i, v_i), $Q_i \in Q$ e v_i é o valor calculado para Q_i em um dado momento.

Definição 3.2. Medida Global de Qualidade: A Medida Global de Qualidade MGQ de uma fonte de dados f, denotada por $f.MGQ$, é um valor obtido a partir da combinação dos valores v_i de cada par $(Q_i, v_i) \in CQ$, com o objetivo de gerar um valor único que indicará o valor global de qualidade de f.

Definição 3.3. Perfil de Qualidade: O Perfil de Qualidade de uma fonte de dados f, denotado por $PQ(f)$, é dado pelo par (CQ, MGQ), onde CQ é o conjunto de valores dos critérios de qualidade do conjunto Q e MGQ é a medida global de qualidade da fonte de dados f.

Definição 3.4. Frequência de Atualização do Perfil de Qualidade: Indica a frequência λ com que $PQ(f)$ será atualizado. Esse valor está diretamente ligado à frequência com que f sofre atualizações. Por exemplo, se a frequência com que f se modifica é diária, então é recomendado que $PQ(f)$ também seja atualizado diariamente.

3.2. Critérios de Qualidade e Métricas de Avaliação

Neste trabalho, utilizamos a estratégia de avaliação contínua da qualidade de uma fonte de dados e geração do seu respectivo Perfil de Qualidade, considerando dois critérios de qualidade: **Completude e Corretude**. Os critérios foram escolhidos por serem dois aspectos pertinentes, que podem ser avaliados independente de uma dada aplicação. Além disso, dados incompletos e incorretos são problemas que ocorrem na grande maioria das fontes de dados, podendo afetar diretamente o uso dos dados. É importante salientar que, o conjunto de critérios de QI utilizados para representar o Perfil de Qualidade é expansível, ou seja, a escolha de um critério bem como a sua usabilidade depende do que se pretende avaliar.

3.2.1. Completude

O critério Completude indica o grau em que os dados de uma fonte são completos para uma determinada tarefa [Wang and Strong 1996]. O valor para esse critério pode ser es-

timado com base nos valores nulos e faltantes, por meio das métricas de Densidade e Cobertura [Naumann and Freytag 2000], respectivamente.

Definição 3.5. Densidade - A densidade D de uma fonte de dados f, denotada por $D(f)$, é dada pelo percentual de valores não nulos contidos em f.

Considere que uma fonte de dados f oferece um conjunto de dados, o qual é descrito pelo conjunto de atributos $A = \{A_1, ..., A_n\}$, onde cada atributo $A_i \in A$ tem um conjunto de valores associados a ele, denotado por $V(A_i)$. Para calcular a densidade $D(f)$ é necessário calcular, primeiramente, a densidade de cada A_i, denotada por $d(A_i)$. A densidade $d(A_i)$ é calculada pela quantidade de valores não nulos contidos em $V(A_i)$, representado por α, sobre a quantidade total de valores contidos em $V(A_i)$, representado por β. Dessa forma, $D(f)$ é dada pelo somatório de todos os valores de $d(A_i) \in A$, dividido pelo número de atributos contidos em A, representado por n (Equação 1).

$$D(f) = \frac{\sum_{i=1}^{n}(d(A_i) = \frac{\alpha}{\beta})}{n} \tag{1}$$

Definição 3.6. Cobertura - A cobertura C de uma fonte de dados f é dada pela quantidade de instâncias disponíveis em f, denotada por $|I|$, com relação a quantidade de instâncias que são esperadas em f, denotada por $|W|$ (Equação 2).

$$C(f) = \frac{|I|}{|W|} \tag{2}$$

Em algumas situações é possível determinar o total de instâncias esperadas em uma fonte de dados. Considerando o nosso exemplo, espera-se que o sensor S faça uma nova coleta de dados a cada hora. Logo, para o período de tempo de 24 horas (1 dia) espera-se que sejam geradas 24 novas instâncias, ou seja, $|W| = 24$. Supondo que no período citado, o sensor deixou de coletar dados por duas horas, teríamos que $|I| = 22$. Se calcularmos a cobertura para S considerando esse período, temos que $C(S) = 0.92$. Em outras palavras, apenas 92% das instâncias que eram esperadas em S estão presentes. Em situações em que não seja possível saber o valor de $|W|$ para f, assume-se que $C(f) = 1$.

Uma vez que foram mensurados o valor para Densidade e Cobertura, o valor da Completude de dados de uma fonte f pode ser dado pela Equação 3.

$$Completude(f) = D(f) * C(f) \tag{3}$$

3.2.2. Corretude

O critério Corretude indica o grau em que os dados contidos em uma fonte de dados estão livres de erros [Wang and Strong 1996, Lóscio et al. 2012]. De maneira geral, é comum encontrar fontes de dados com erros, seja por falha humana ou por falhas na coleta e na geração dos dados. No entanto, avaliar a Corretude de uma fonte de dados pode ser uma tarefa difícil, uma vez que a identificação de dados incorretos pode depender de um amplo conhecimento do domínio e da presença de um especialista.

Neste trabalho, a fim de avaliar o critério de Corretude, propomos o uso de regras de validação, as quais são definidas de acordo com o domínio dos dados e podem ser usadas para uma avaliação automática da Corretude. Tais regras podem ser definidas

por um especialista do domínio. Sendo assim, assumimos que cada fonte de dados f está associada a um conjunto de *Regras de Validação* $R = \{R_1, ..., R_n\}$ que descrevem restrições do domínio de dados de f.

Definição 3.7. Regra de Validação - Cada regra de validação $R_i \in R$ é descrita por uma ou mais expressões de validação ($Exp_1, ..., Exp_m$) conectadas por operadores lógicos (&&, ||, !). Cada expressão de validação Exp_j define uma restrição do domínio de dados de f, tal que $Exp_j := exp_{j1}\ \varphi\ exp_{j2}$, onde exp_{j1} é um atributo $A_i \in f.A$, exp_{j2} pode ser um atributo $A_k \in f.A$ ou um literal e φ é um operador relacional ($<$, $>$, $\leq$, $\geq$, $=$, $\neq$).

Para ilustrar a definição das regras de validação, considere o exemplo da Seção 2 e a seguinte restrição do domínio de dados meteorológicos: *a temperatura mínima deve ser menor que a temperatura média, que por sua vez deve ser menor ou igual à temperatura máxima*.

Nesse caso, considerando que os dados coletados pelo sensor S são descritos pelo conjunto de atributos $A = \{data\_medicao,\ temp\_minima,\ temp\_media,\ temp\_maxima\}$, poderíamos usar a seguinte regra $R_1 = \{temp\_minima < temp\_media\ \&\&\ temp\_media \leq temp\_maxima\}$ como forma de validação da restrição descrita acima. Uma vez que o conjunto de regras R foi definido, é possível avaliar a corretude da fonte de dados. De forma simplificada, uma instância que possuir valores que não atendam às regras estabelecidas é sinalizada como uma instância incorreta.

O valor para o critério de Corretude de uma fonte de dados f pode ser obtido pela Equação 4. Primeiro, a partir da verificação das regras de validação, calcula-se o número de instâncias incorretas da fonte f (δ). Em seguida, esse valor é dividido pelo número total de instâncias de f (β) e o resultado é subtraído de um (1) para que seja encontrada a porcentagem de instâncias corretas da fonte f.

$$Corretude(f) = 1 - \left(\frac{\delta}{\beta}\right) \tag{4}$$

4. Geração do Perfil de Qualidade

Nesta seção, descrevemos o processo de geração do Perfil de Qualidade de uma fonte de dados f, ilustrado pela Figura 1 e descrito a seguir. A geração do Perfil de Qualidade segue uma estratégia de avaliação contínua da qualidade e deverá ser atualizado de acordo com uma frequência de atualização previamente definida (λ).

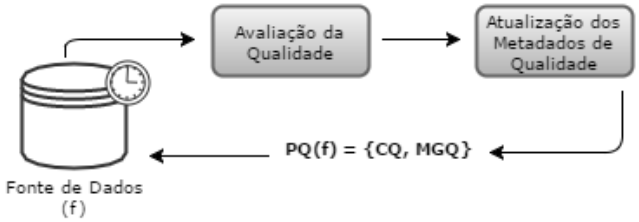


Figura 1. Visão Geral do Processo

A primeira etapa do processo de geração do Perfil de Qualidade é a **Avaliação da Qualidade**. Nessa etapa, os critérios de qualidade $Q_i \in Q$ serão avaliados conforme descrito na Seção 3.2. É importante destacar que, no momento da avaliação da qualidade,

ao invés de considerar todo o conjunto de instâncias presentes na fonte, consideramos apenas uma amostra. Tal ação permite que haja uma redução no custo computacional para avaliar a qualidade da fonte, uma vez que o volume dos dados a serem avaliados pode ser muito grande. Especificamente, serão avaliadas apenas as novas instâncias que foram adicionadas na fonte de dados no intervalo de tempo compreendido entre o instante de tempo da última avaliação de qualidade da fonte e o instante de tempo da avaliação atual.

Porém, como discutido anteriormente, considerar apenas um amostra isolada dos dados pode não ser suficiente para avaliar com precisão a qualidade de uma fonte. Dessa forma, para que os valores dos critérios de qualidade sejam o mais próximo possível dos valores reais, consideramos também os resultados de avaliações de qualidade ocorridas em instantes anteriores.

A segunda etapa do processo de geração do Perfil de Qualidade é a **Atualização dos Metadados de Qualidade**. O primeiro passo dessa etapa consiste em calcular os valores de v_i para cada um dos elementos do conjunto CQ de $PQ(f)$. Sendo assim, para cada critério $Q_i \in Q$, seu novo valor, denotado de $Q_i.v_{i_{new}}$, será calculado como mostra a Equação 5.

$$Q_i.v_{i_{new}} = w * \Delta(Q_i) + (1 - w) * (Q_i.v_i) \tag{5}$$

Onde, $Q_i.v_{i_{new}}$ é o novo valor do critério Q_i que será atualizado em $CQ \in PQ(f)$. Este valor é calculado com base no valor obtido na avaliação realizada na etapa anterior, denotado por $\Delta(Q_i)$, e pelo valor corrente do critério Q_i, denotado por $Q_i.v_i$. Nesta equação, w corresponde ao peso que $\Delta(Q_i)$ assumirá no cálculo. O valor de w equivale à proporção de instâncias que foram avaliadas (pode ser obtido pelo total de instâncias que foram avaliadas dividido pelo total de instâncias contidas na fonte).

Após a atualização dos valores de qualidade da fonte de dados, o novo valor de $f.MGQ$ deve ser calculado (Equação 6).

$$MGQ_{new} = \sum_{i=1}^{n} w_i * Q_i.v_i \tag{6}$$

Onde, $n = |Q|$ e w_i corresponde ao peso atribuído ao critério Q_i. É importante salientar que a soma dos pesos não pode ser maior que 1. No contexto da MGQ, o peso atribuído a um critério indica a relevância que ele exerce sobre a avaliação geral da fonte de dados.

Uma vez que foram realizados os cálculos correspondentes à etapa de atualização dos metadados de qualidade, podemos dizer que $PQ(f)$ foi devidamente atualizado. A Figura 2(a) apresenta a estrutura do Perfil de Qualidade e a Figura 2(b) mostra um exemplo do Perfil de Qualidade representado no formato JSON.

5. Experimentos

Como forma de avaliar nossa proposta, mais especificamente a avaliação contínua da qualidade das fontes de dados, realizamos alguns experimentos que serão descritos a seguir. Todos os experimentos foram realizados em um computador com processador Intel Core i5 2.40GHz e 4GB de memória. As etapas do processo de geração do Perfil de Qualidade, proposta na Seção 4, foram implementadas na linguagem Java.

Elementos	Descrição
URL	URL de acesso a fonte de dados.
Última atualização da fonte de dados	Data que ocorreu a modificação mais recente na fonte de dados.
Última modificação do Perfil de Qualidade	Data que o Perfil de Qualidade foi modificado.
Volume de Dados	Quantidade de instâncias contidas na fonte de dados.
Critérios de QI	Informações de Qualidade para cada critério avaliado.
Medida Global	Valor de Qualidade Global da fonte de dados

(a) Estrutura do Perfil de Qualidade


```
{
   "url_fonte": "http://fontededados.com.br/acesso",
   "ultima_atualizacao": "20/05/2016 17:00:01",
   "geracao_perfil": "20/05/2016 17:15:06",
   "volume_dados": 12.395,
   "criterios_qi": [{
        "nome_criterio": "Completude",
        "valor_criterio": 0.89
   }, {
        "nome_criterio": "Corretude",
        "valor_criterio": 0.95
   }],
   "medida_global": 0.92
}
```


(b) Exemplo em JSON

Figura 2. Ilustração do Perfil de Qualidade

Em nossos experimentos, consideramos três estratégias de avaliação da qualidade das fontes: (i) **avaliação pontual**, a estratégia que avalia a qualidade de uma fonte com base em uma amostra de dados, de forma semelhante à utilizada em Xian et al. [Xian et al. 2009]. Especificamente, a amostra retirada para avaliação corresponde às novas instâncias incluídas na fonte após a última avaliação de qualidade da fonte; (ii) **avaliação ideal**, a estratégia que reavalia toda a fonte quando novas instâncias são recebidas e (iii) **avaliação contínua**, a estratégia proposta nesse trabalho (descrita na Seção 4).

Os experimentos objetivaram validar as seguintes hipóteses: H1 - A estratégia de avaliação contínua, quando comparada com a estratégia de pontual, fornece resultados mais próximos da estratégia ideal. H2 - Quando comparada com a avaliação ideal, há pouca perda de precisão nos valores dos critérios de qualidade calculados pela estratégia de avaliação contínua. H3 - A estratégia de avaliação contínua proporciona uma redução no tempo de execução da avaliação de qualidade quando comparada com a avaliação ideal.

5.1. Fontes de Dados

O domínio de dados das fontes utilizadas para os os experimentos foi o Meteorológico. Utilizamos duas (2) fontes de dados de duas (2) instituições que monitoram as condições climáticas da cidade do Recife (ITEP e APAC). A escolha por essas fontes se deu justamente por elas serem dinâmicas, uma vez que atualizações de inserção são frequentes. Dessa forma, o volume de dados produzido pode crescer consideravelmente ao longo do tempo.

Entretanto, como essas instituições não oferecem uma API ou um serviço que permita o acesso aos dados em tempo real, não foi possível realizar o monitoramento nem a coleta contínua e automática dos dados. Dessa forma, fizemos uma solicitação e cada instituição nos forneceu arquivos com os dados em formato CSV. Os dados utilizados em nossos experimentos foram coletados no período de 01/01/2013 a 31/12/2014, totalizando 12.255 instâncias para a fonte da APAC e 12.265 instâncias para a fonte do ITEP.

Os dados dessas fontes foram coletados por sensores instalados em uma estação meteorológica (EM). As EMs possuem diversos tipos de sensores, no entanto, neste trabalho, avaliamos apenas os dados oriundos dos sensores de temperatura. Para avaliar a qualidade dessas fontes, utilizamos os critérios de Corretude e Completude, descritos na Seção 3.2. Considerando o domínio das fontes de dados avaliadas, o conjunto de regras de validação para o critério de Corretude (Seção 3.2.2) foi implementado a partir de algumas

validações básicas, que compreendem três regras específicas para o domínio em questão, propostas no trabalho de Baba et al. [Baba et al. 2014].

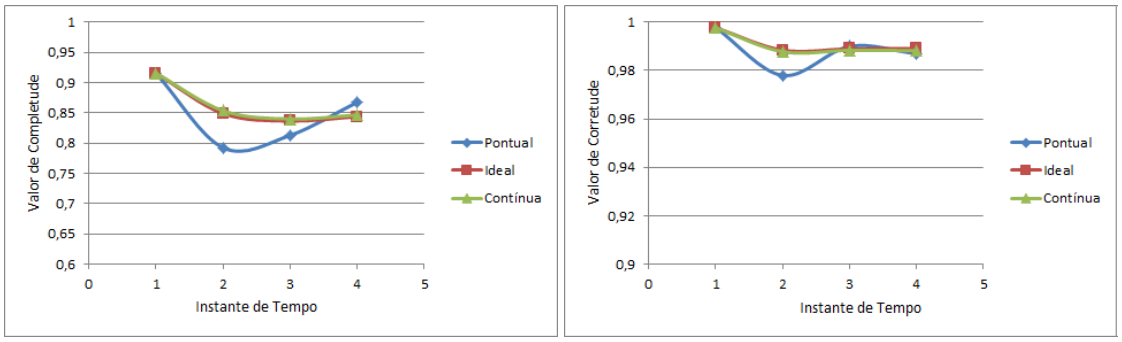
Como não foi possível monitorar as fontes, dividimos, proporcionalmente, os conjuntos de dados em lotes. Em nossos experimentos, cada lote foi utilizado para simular um instante de tempo em que a fonte sofreu uma atualização de inserção. A quantidade de instâncias inseridas na fonte de dados em cada instante pode ser vista na Tabela 1. É importante destacar que, nesse caso, a fonte seria o banco de dados que armazena os dados coletados pelo sensor e não o sensor propriamente dito.

Tabela 1. Lotes de Dados

Lotes de Dados					
Fonte de Dados	t_1	t_2	t_3	t_4	**Volume Total**
APAC	3.098	3.040	3.240	2.977	12.255
ITEP	3.086	3.070	3.138	2.971	12.265

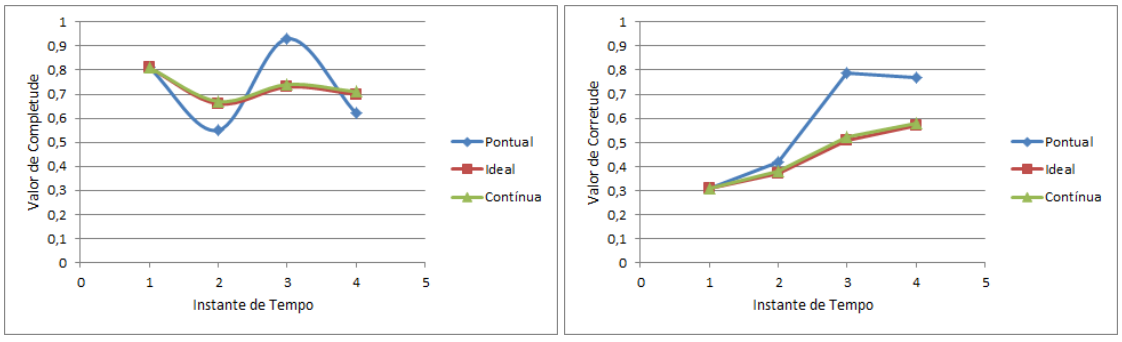
5.2. Resultados

Para cada instante de tempo (t_1, t_2, t_3 e t_4), realizamos a avaliação dos critérios de qualidade utilizando as três (3) estratégias mencionadas acima e foram obtidos os seguintes resultados (Figuras 3 e 4).



(a) Completude (b) Corretude

Figura 3. Resultado das Avaliações - Fonte APAC

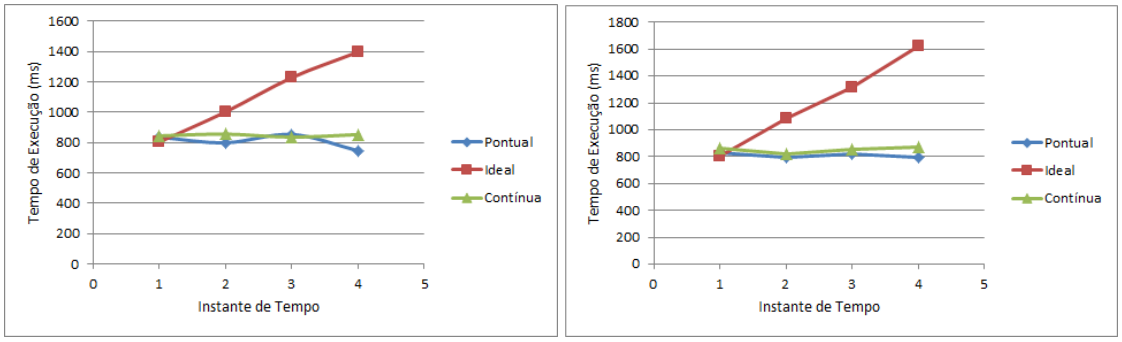


(a) Completude (b) Corretude

Figura 4. Resultado das Avaliações - Fonte ITEP

Durante a execução dos experimentos, calculamos o tempo gasto para realizar a avaliação dos critérios utilizando as três estratégias mencionadas acima. Ao final, foram

obtidos os resultados apresentados na Figura 5. É importante salientar que para cada instante de tempo executamos cada estratégia cinco (5) vezes para realizar uma média ponderada dos tempos obtidos em cada uma delas.



(a) APAC (b) ITEP

Figura 5. Tempo de Execução (ms) da Avaliação da Qualidade

5.3. Discussões

Fazendo uma análise dos resultados obtidos com os experimentos, podemos concluir que, ao longo do tempo, a qualidade de uma fonte de dados pode mudar em decorrência das inserções de novos dados. Dessa forma, ficou evidente a importância de uma avaliação contínua da qualidade em fontes de dados dinâmicas.

Com relação à estratégia de avaliação pontual, em nossos experimentos, a mesma se mostrou ineficiente. Tanto na fonte de dados da APAC (Figura 3), quanto na do ITEP (Figura 4), os valores atribuídos aos critérios de qualidade, em diferentes instantes de tempo, tiveram valores distantes do esperado quando comparados com a estratégia ideal. Isso ocorre porque, a cada instante de tempo, essa estratégia considera apenas as novas instâncias para calcular os valores dos critérios. Dessa forma, a avaliação da qualidade não consegue refletir a evolução da qualidade da fonte ao longo do tempo.

Por outro lado, a estratégia de avaliação contínua apresentou bons resultados em ambas as fontes de dados avaliadas. Em todos os instantes de tempo, os valores atribuídos aos critérios de qualidade tiveram valores bem aproximados do esperado, quando comparada com a estratégia ideal. Isso ocorre porque, apesar de avaliar apenas as novas instâncias, o cálculo dos valores dos critérios de qualidade também considera os resultados das avaliações anteriores. Dessa forma, o resultado obtido com a estratégia de avaliação contínua consegue refletir, com mais precisão, a evolução da qualidade da fonte ao longo do tempo, mesmo sem ter que reavaliá-la por completo.

Com relação ao custo computacional, das três estratégias, a avaliação pontual apresentou um menor custo computacional. No entanto, como discutido, os resultados obtidos com essa estratégia não conseguem refletir com precisão a qualidade da fonte de dados. Em comparação com a estratégia pontual, a estratégia contínua apresenta pouca diferença em termos de custo computacional. Por exemplo, no instante t_4 houve um aumento de 12% (APAC) e 9% (ITEP) (Figura 5). É possível observar que a estratégia da avaliação contínua se mantém estável, enquanto na estratégia ideal a tendência é crescente ao longo do tempo. Esse é um bom indicativo, pois mostra que, optando pela estratégia de avaliação contínua, é possível ter uma redução considerável no tempo de execução da

avaliação. Por exemplo, quando comparamos o instante t_4, onde o volume de dados acumulado nas fontes é maior, há uma redução no tempo de execução da avaliação de 63% (APAC) e 86% (ITEP) sem que haja uma perda de precisão nos valores dos critérios.

6. Trabalhos Relacionados

Em Mihaila et al. [Mihaila et al. 2000], os autores propõem o uso de metadados de qualidade para auxiliar na tarefa de seleção de fontes de dados. O enfoque principal desse trabalho consiste em definir um modelo no qual os metadados possam ser representados e posteriormente consultados. Diferentemente, nosso trabalho propõe um processo para geração desses metadados. Especificamos e definimos como os critérios de qualidade serão mensurados e avaliados ao longo do tempo, o que faltou no trabalho citado acima, uma vez que os autores assumem que os proprietários das fontes fornecem as informações de qualidade.

Com relação à avaliação da qualidade em fontes de dados, a literatura oferece vários trabalhos que utilizam informações de qualidade associadas às fontes para diversos fins. Por exemplo, Xian et al. [Xian et al. 2009] utilizam informações de qualidade para auxiliar a selecionar as melhores fontes de dados para um Sistema de Integração de Dados. Lóscio et al. [Lóscio et al. 2012] propõem o uso de informações de qualidade para auxiliar o desenvolvedor de uma aplicação específica a identificar fontes de dados mais relevantes para sua aplicação.

Apesar de terem um processo de avaliação bem definido, os trabalhos citados realizam avaliações pontuais nas fontes de dados. Por exemplo, o trabalho de Xian et al. considera apenas uma amostra dos dados para avaliar as fontes. A proposta de Lóscio et al. [Lóscio et al. 2012] considera requisitos da aplicação para avaliar os critérios de qualidade, o que torna a avaliação muito específica. O problema de uma avaliação específica é que os resultados não podem ser reutilizados em outros cenários. Diferentemente, nosso trabalho propõe uma avaliação para um contexto geral de utilização da fonte de dados.

Dentre os poucos trabalhos que abordam a problemática da avaliação da qualidade em fontes de dados dinâmicas destaca-se o de Rekatsinas et al. [Rekatsinas et al. 2014]. O trabalho de Rekatsinas et al. [Rekatsinas et al. 2014] trata a questão da dinamicidade das fontes no contexto de Integração de Dados. Os autores avaliam as fontes de dados de maneira contínua com base na sua frequência de modificação. Essa avaliação é usada para estimar o comportamento da fonte de dados em um tempo futuro e descobrir um melhor conjunto de fontes a serem integradas ao longo do tempo.

7. Conclusão

Neste artigo, propomos um processo para geração de um Perfil de Qualidade para fontes de dados dinâmicas. O processo é composto pelas etapas da avaliação da qualidade e geração dos metadados de qualidade. Um dos diferenciais da nossa proposta é que, na etapa da avaliação da qualidade, a natureza dinâmica das fontes é considerada.

Nossos experimentos comprovaram a importância de uma avaliação contínua nas fontes de dados e indicaram que a estratégia proposta é capaz de alcançar resultados muito próximos do esperado. Outro ganho da estratégia de avaliação contínua da qualidade é com relação ao custo computacional. Ao longo do tempo, a tendência é que o volume de dados cresça e o custo para processá-los seja alto. Os resultados indicaram que utilizando a estratégia proposta é possível ter uma redução considerável desses custos.

Como indicações para trabalhos futuros, pretendemos enriquecer as informações contidas no Perfil de Qualidade, de forma a proporcionar uma avaliação de qualidade mais ampla, com critérios relacionados à qualidade do serviço das fontes [Dustdar et al. 2012]. Também pretendemos realizar novos experimentos utilizando fontes de dados de outros domínios.

Referências

Baba, R. K., Vaz, M. S. M. G., and Costa, J. (2014). Correção de dados agrometeorológicos utilizando métodos estatísticos. *Revista Brasileira de Meteorologia*, 29(4).

Dong, X. L., Saha, B., and Srivastava, D. (2013). Less is more: selecting sources wisely for integration. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 37–48. VLDB Endowment.

Duquennoy, S., Grimaud, G., and Vandewalle, J. J. (2009). The web of things: Interconnecting devices with high usability and performance. In *Embedded Software and Systems, 2009. ICESS '09. International Conference on*, pages 323–330.

Dustdar, S., Pichler, R., Savenkov, V., and Truong, H.-L. (2012). Quality-aware service-oriented data integration: Requirements, state of the art and open challenges. *SIGMOD Rec.*, 41(1):11–19.

Lóscio, B. F., Batista, M. C. M., Souza, D., and Salgado, A. C. (2012). Using information quality for the identification of relevant web data sources: A proposal. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, IIWAS '12, pages 36–44, New York, NY, USA. ACM.

Lóscio, B. F., Burle, C., and Calegari, N. (2016). Data on the web best practices.

Malaverri, J. E. G., Santanche, A., and Medeiros, C. B. (2014). A provenance-based approach to evaluate data quality in escience. *Int. J. Metadata Semant. Ontologies*, 9(1):15–28.

Mihaila, G. A., Raschid, L., and Vidal, M. (2000). Using quality of data metadata for source selection and ranking. In *Proceedings of the Third International Workshop on the Web and Databases, WebDB*, pages 93–98.

Naumann, F. and Freytag, J. C. (2000). Completeness of information sources. Technical report, Humboldt University of Berlin.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211–218.

Rekatsinas, T., Dong, X. L., and Srivastava, D. (2014). Characterizing and selecting fresh data sources. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD'14, pages 919–930, Snowbird, Utah, USA. ACM.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33.

Xian, X.-F., Zhao, P.-P., Fang, W., Xin, J., and Cui, Z.-M. (2009). Quality-based data source selection for web-scale deep web data integration. In *2009 International Conference on Machine Learning and Cybernetics*, volume 1, pages 427–432.