Pós-Graduação em Ciência da Computação

Everaldo Costa Silva Neto

UM PERFIL DE QUALIDADE PARA FONTES DE DADOS DINÂMICAS

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
<www.cin.ufpe.br/~posgraduacao>

RECIFE
2016

Everaldo Costa Silva Neto

UM PERFIL DE QUALIDADE PARA FONTES DE DADOS DINÂMICAS

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: *Ana Carolina Brandão Salgado*
Co-Orientadora: *Bernadette Farias Lóscio*

RECIFE
2016

Catalogação na fonte
Bibliotecária Joana D'Arc Leão Salvador CRB 4-572

S586p Silva Neto, Everaldo Costa.
Um perfil de qualidade para fontes de dados dinâmicas / Everaldo Costa Silva Neto. – 2016.
92 f.: fig., tab., quad.

Orientadora: Ana Carolina Brandão Salgado.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIN. Ciência da Computação, Recife, 2016.
Inclui referências.

1. Ciência da computação. 2. Banco de dados. 3. Qualidade da informação. I. Salgado, Ana Carolina Brandão (Orientadora). II. Titulo.

004 CDD (22. ed.) UFPE-MEI 2016-122

Everaldo Costa Silva Neto

Um Perfil de Qualidade para Fontes de Dados Dinâmicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 24/08/2016

BANCA EXAMINADORA

_____

Prof. Dr. Fernando da Fonseca de Souza
Centro de Informática / UFPE

_____

Profa. Dra. Maria da Conceição Moraes Batista
Departamento de Estatística e Informática/UFRPE

_____

Profa. Dra. Ana Carolina Brandão Salgado
Centro de Informática /UFPE
(**Orientadora**)

Dedico este trabalho a todos que me apoiaram e se fizeram presente, mesmo que de longe, no decorrer dessa longa jornada.

Agradecimentos

Gratidão é a palavra que resume o meu sentimento por ter conseguido chegar ao fim desta jornada. Foi uma experiência incrível e desafiadora, desde os primeiros dias até os últimos, e lembrando toda trajetória eu só consigo ser grato.

Esse tempo foi de bastante aprendizado, intelectual e emocional. A montanha russa de emoções vivida nesse período me mostrou que eu não era uma fortaleza como imaginava, pelo contrário... Por muitas vezes eu tive que me reinventar e superar minhas limitações. A lição que eu tiro é que tudo é possível quando temos um objetivo.

Nesses dois últimos anos várias pessoas foram fundamentais para que eu tivesse conseguido concluir essa etapa, e gostaria de agradecer, de maneira especial, a cada um.

Agradeço a Deus por ter me permitido viver tudo isso e ter me abençoado, suprindo e providenciando tudo aquilo que precisava. Tenho certeza que se não fosse a sua boa mão me guiando nada disso seria possível.

Agradeço a minha família, em especial aos meus pais, Emerson e Geórgia, por terem me apoiado e serem meus maiores incentivadores, desde a época que saí de casa para fazer a graduação até quando tomei a decisão de mudar para o Recife e vir fazer o mestrado. Obrigado pelas palavras de ânimo e por me cercarem com orações. Também não poderia deixar de agradecer a minha tia, Maria Helena, por ter me recebido e acolhido em sua casa durante todo esse tempo do mestrado. Esse apoio foi fundamental.

Agradeço as minhas orientadoras, Carol e Berna, por terem me dado a oportunidade de trabalhar e aprender junto delas. Vocês são pessoas incríveis e admiráveis, obrigado por sempre serem solícitas, por terem acreditado em mim, pelos conselhos, pela paciência, e principalmente, por ter conduzido este trabalho. A visão e o senso crítico de vocês foram sempre o meu maior incentivo na busca por fazer algo melhor.

Nesse tempo eu também fiz novas amizades que quero manter por muito tempo. Agradeço ao Edson, Fran e Conrado, essa família que foi tão importante pra mim no primeiro ano aqui em Recife, obrigado por compartilhar dos passeios e churrascos nos fins de semana. Agradeço ao pessoal da "Toca dos Dados"pelos momentos de debates e aprendizados em conjunto, em especial gostaria de agradecer a Diego, Marcelo, Priscilla e Gabi.

Agradeço ao Diego por ter sido sempre tão prestativo quando precisei. Agradeço ao Marcelo pelos conselhos e pelas incontáveis ajudas que me deu no começo da pesquisa. Agradeço a Priscilla, companheira diária de laboratório e de testes (divinos, hahaha), por ter acompanhado cada descoberta e ouvido (alguns) lamentos nesse último ano.

Agradeço a Gabi, minha amiga e companheira nessa trajetória, estivemos juntos, literalmente, desde o primeiro até o último dia dessa experiência. Foi muito bom poder compartilhar das alegrias e dissabores, desde as disciplinas até a fase final da pesquisa. Pode ter certeza que

vou lembrar sempre com carinho dos momentos que partilhamos dentro e fora do CIn, até dos arranca-rabos no projeto de Integração de Dados (hahaha). Obrigado por tudo, você foi muito importante, eu espero que possamos manter nossa amizade mesmo que a vida nos leve para caminhos distantes.

Agradeço a Vilma, pela torcida, por sempre me encorajar, e principalmente, pela paciência em me ouvir e aconselhar quando eu precisava. Tenho certeza que os bons pensamentos e as orações foram importantes.

Agradeço ao professor Fernando e a professora Maria da Conceição (Ceça) por terem aceito o convite para participar da banca examinadora deste trabalho. Vocês são referência para mim. Fernando, com seu jeito peculiar de lecionar (e fazer provas, hahaha), onde tive o prazer de tê-lo como professor em duas disciplinas do Mestrado e Ceça, por ter sido, quando resolvi investigar a área de Qualidade da Informação, uma das primeiras referências que li sobre o tema.

Por fim, gostaria de agradecer a todos que torceram e contribuíram, direta ou indiretamente, para realização deste trabalho. Obrigado!!!!!! :)

Todos nós deveríamos nos preparar para o futuro, aprendendo coisas que ainda não sabemos, desaprendendo coisas que sabemos, mas que não deveríamos mais saber, porque são de alguma forma inúteis, e reaprendendo coisas que já soubemos e que voltaram a ser úteis.

—SILVIO MEIRA

Resumo

Atualmente, um massivo volume de dados tem sido produzido pelos mais variados tipos de fontes de dados. Apesar da crescente facilidade de acesso a esses dados, identificar quais fontes de dados são mais adequadas para um determinado uso é um grande desafio. Isso ocorre devido ao grande número de fontes de dados disponíveis e, principalmente, devido à ausência de informações sobre a qualidade dos dados. Nesse contexto, a literatura oferece diversos trabalhos que abordam o uso de critérios de Qualidade da Informação (QI) para avaliar fontes de dados e solucionar esse desafio. No entanto, poucos trabalhos consideram o aspecto dinâmico das fontes na etapa da avaliação da qualidade. Nesta dissertação, abordamos o problema de avaliação da qualidade em fontes de dados dinâmicas, ou seja, fontes de dados cujo conteúdo pode sofrer modificações com alta frequência. Como contribuição, propomos uma estratégia onde os critérios de QI são avaliados de forma contínua, com o objetivo de acompanhar a evolução das fontes de dados ao longo do tempo. Além disso, propomos a criação de um Perfil de Qualidade, que consiste de um conjunto de metadados sobre a qualidade de uma fonte, onde seu uso pode ser aplicado para diversos fins, inclusive no processo de seleção de fontes de dados. O Perfil de Qualidade proposto é atualizado periodicamente de acordo com os resultados obtidos pela avaliação contínua da qualidade. Dessa forma, é possível refletir o aspecto dinâmico das fontes. Para avaliar os resultados deste trabalho, mais especificamente a estratégia de avaliação contínua da qualidade, utilizamos fontes de dados do domínio Meteorológico. Os experimentos realizados demonstraram que a estratégia de avaliação proposta produz resultados satisfatórios.

Palavras-chave: Fontes de Dados Dinâmicas. Qualidade da Informação. Avaliação da Qualidade. Metadados de Qualidade. Perfil de Qualidade

Abstract

Nowadays, a massive data volume 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. Several works in the literature use Information Quality as a mean of solving this problem, however, only few works employ a continuous strategy. In this work, we address the problem of performing assessment continuously, looking to dynamic data sources. We also propose the creation of a data source Quality Profile, which consists of a set of metadata about the data source's quality and may be used to help the selection of data sources. To reflect the real quality values of a data source, we propose a continuous updating of the Quality Profile, according to the data source's refresh rate. In order to evaluate our proposal, we carried out some experiments with meteorological data provided by institutions that monitor weather conditions of Recife. The experimental results have demonstrated that our strategy produces more satisfactory results than others, regarding the trade off between performance and accuracy.

Keywords: Data Sources Dynamics. Information Quality. Quality Assessment. Quality Metadata. Quality Profile

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Acrônimos

Sumário

1 Introdução

Este capítulo fornece uma visão geral desta pesquisa e apresenta o contexto no qual este trabalho está inserido. Primeiramente, mostramos a motivação do estudo e a caracterização do problema abordado. Depois, elencamos os objetivos pretendidos, seguindo por uma discussão sobre as contribuições esperadas. Por fim, apresentamos a descrição da estrutura desta dissertação.

1.1 Motivação

Nos últimos tempos, um massivo volume 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* (LOSCIO; BURLE; CALEGARI, 2016) e o paradigma de WoT (*Web of Things*) (DUQUENNOY; GRIMAUD; VANDEWALLE, 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, principalmente, devido à ausência de informações sobre a qualidade dos dados.

Por exemplo, a *Web* é um ambiente heterogêneo, que abriga fontes com domínios e formatos de dados variados. Além disso, por ser um ambiente flexível e não possuir mecanismos que verifiquem ou analisem os dados que são armazenados e disponibilizados nas fontes, é comum encontrar fontes de dados com baixa qualidade (REKATSINAS et al., 2015). Nesse contexto, esse desafio torna-se maior, e a presença de informações que descrevam a qualidade das fontes de dados se apresenta como um facilitador.

Considere o exemplo a seguir para ilustrar uma situação que ocorre no mundo real. Suponha que um usuário esteja interessado em utilizar dados que descrevam uma determinada entidade. Imagine que múltiplas fontes podem fornecer dados sobre a entidade de interesse. O custo para analisar e integrar dados de múltiplas fontes pode ser alto (XIAN et al., 2009; REKATSINAS; DONG; SRIVASTAVA, 2014). Uma alternativa para minimizar esses custos é

selecionar, dentre um conjunto de fontes de dados, aquelas que são melhores e/ou mais adequadas para o uso em questão.

Em suma, o processo de seleção de fontes pode ser descrito da seguinte forma: (i) identificar um conjunto de fontes de dados que são candidatas a uso; (ii) avaliar a qualidade dessas fontes; (iii) com base na qualidade, selecionar aquela(s) que mais se adequa(m) à sua necessidade. Normalmente, a qualidade de uma fonte de dados é dada com base em um conjunto de características, que vão desde aspectos ligados à proveniência (MALAVERRI; SANTANCHE; MEDEIROS, 2014) e qualidade dos dados (LOSCIO et al., 2012; REKATSINAS et al., 2015), até à qualidade do serviço que é provido pelas fontes (DUSTDAR et al., 2012).

De uma maneira geral, as etapas descritas acima tendem ser custosas para o usuário, uma vez que as fontes de dados não fornecem informações acerca de sua qualidade, ficando a cargo do usuário fazer essa análise de forma manual. Analisar manualmente as fontes de dados pode ser uma tarefa desafiadora, devido a vários fatores como: o número de fontes consideradas pode ser alto e muitas vezes as fontes de dados podem ser desconhecidas pelo usuário. Outros fatores também podem contribuir, por exemplo, as fontes de dados podem ser heterogêneas, ou seja, ter seus dados representados de maneira distinta (hierárquico, tabular, grafo) ou até mesmo formatos distintos para uma mesma representação (ex.: RDF e OWL para representar grafos).

Nesse sentido, 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; LOSCIO et al., 2012; DONG; SAHA; SRIVASTAVA, 2012). Em geral, esses trabalhos propõem o uso de critérios de Qualidade da Informação (QI) (WANG; STRONG, 1996), tais como disponibilidade, corretude, completude dos dados, entre outros, 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.

Nesta dissertação, 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; DONG; SRIVASTAVA, 2014). 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.

A fim de considerar o aspecto dinâmico das fontes de dados 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 em avaliações anteriores.

Além disso, este trabalho tem como objetivo especificar um processo para geração de um

Perfil de Qualidade para fontes de dados dinâmicas. Especificamente, tal processo engloba duas etapas (i) avaliação da qualidade das fontes de dados e (ii) geração dos metadados de qualidade.

Na primeira etapa do processo, as fontes de dados são avaliadas por meio de critérios de QI, utilizando a estratégia de avaliação contínua. Em seguida, informações de qualidade da fonte são geradas e reunidas em um Perfil de Qualidade.

O Perfil de Qualidade consiste de um conjunto de metadados que poderão ser disponibilizados juntamente com a fonte de dados e poderão ser utilizados para facilitar o processo de seleção de fontes de dados, por exemplo. Por estarmos tratando com fontes de dados dinâmicas, 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 sua qualidade.

1.2 Caracterização do Problema

Nesta seção, apresentamos um exemplo que ilustra o problema de avaliação da qualidade de fontes dinâmicas. Uma fonte de dados é considerada dinâmica quando o seu contéudo sofre modificações com alta frequência. É importante destacar que, nesta dissertação apenas as inclusões de novos dados são tratadas como modificações 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 Ambitente deseja desenvolver uma aplicação para oferecer análises sobre os 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 corretamente 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 atualização 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, 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 acima, 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. (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 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 das fontes. Dessa forma, dependendo da amostra escolhida, a avaliação da qualidade poderá ser prejudicada.

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

Nossa hipótese é que, por meio de uma avaliação contínua, a informação de qualidade oferecida seja mais próxima da realidade do que considerando uma avaliação pontual. Também é esperado que ao longo do tempo a avaliação da qualidade realizada de maneira contínua, seguindo a estratégia proposta neste trabalho, quando comparada com a solução ideal (reavaliando a fonte por completo todas as vezes que uma modificação ocorre), tenha um baixo índice com relação à perda de precisão dos resultados obtidos e um alto índice com relação à redução do custo computacional citado anteriormente.

1.3 Objetivos

Esta dissertação tem como principal objetivo a geração de um Perfil de Qualidade para fontes de dados dinâmicas, o qual será gerado e atualizado de acordo com uma avaliação contínua da qualidade das fontes de dados. Estabelecemos alguns objetivos específicos para alcançar os objetivos gerais deste trabalho:

- Especificação de uma estratégia de avaliação contínua, para tratar o problema da avaliação da qualidade das fontes de dados dinâmicas;
- Definição de métricas para cada critério de QI selecionado para compor o Perfil de Qualidade;
- Especificação de um processo para geração do Perfil de Qualidade;
- Implementação de um protótipo para automatizar o processo de geração do Perfil de Qualidade;
- Realização de experimentos para avaliação da estratégia proposta.

A principal contribuição desta dissertação é a especificação de um processo para geração de um Perfil de Qualidade para as fontes de dados dinâmicas, uma vez que no levantamento do

estado da arte foi possível identificar diversos trabalhos que utilizam informações de qualidade das fontes para diversos fins, justificando assim a utilidade do Perfil de Qualidade. De acordo com o levantamento realizado na literatura, não foi encontrado nenhum trabalho que descreve todo o processo para geração de metadados de qualidade da maneira que propomos. Como parte integrante do processo, outra contribuição é considerar o aspecto dinâmico das fontes de dados na etapa de avaliação, propondo uma estratégia de avaliação contínua.

1.4 Estrutura da Dissertação

Além do presente capítulo, este trabalho está dividido como segue:

- O Capítulo 2 introduz os fundamentos acerca dos principais conceitos básicos para o desenvolvimento desta pesquisa;
- O Capítulo 3 descreve e discute os trabalhos relacionados;
- O Capítulo 4 especifica o processo proposto para geração de um Perfil de Qualidade para as fontes de dados dinâmicas;
- O Capítulo 5 discute os experimentos que foram realizados para validar nossas hipóteses e apresenta os resultados obtidos por meio de tais experimentos;
- O Capítulo 6 apresenta algumas considerações sobre esta dissertação, as contribuições alcançadas, e discute algumas indicações para trabalhos futuros.

2

Fundamentos

Neste capítulo iremos apresentar os principais conceitos acerca dos tópicos de pesquisa que norteiam esta dissertação. Inicialmente discutiremos o conceito de Qualidade da Informação, em seguida iremos abordar o problema de Seleção de Fontes de Dados e por fim introduziremos o conceito de *Data Profiling*.

2.1 Qualidade da Informação

Atualmente estamos vivendo na era da informação. Milhares de dados são produzidos a cada instante e são utilizados para gerar informações. Muitas vezes essas informações são utilizadas para tomar diversas decisões, que podem ser desde as mais simples até as mais complexas. Independente da aplicação da informação, existe uma premissa geral: quanto mais valor a informação agrega, melhor para o seu propósito. Logo, pode-se concluir que o valor da informação está intrinsicamente ligado à sua qualidade (CAI; ZHU, 2015).

QI como um conceito pode ser comumente definida como um conjunto de critérios ou dimensões utilizados para indicar o grau de qualidade geral de uma informação obtida por um sistema (BATISTA, 2008). É classicamente definida pelo termo *fitness for use*, ou adequação ao uso (WANG; STRONG, 1996), o que leva a considerar que a informação é apropriada se atende a um conjunto de requisitos estabelecidos, seja por um usuário ou por um conjunto de normas (SARINHO, 2014).

A literatura oferece uma grande quantidade de trabalhos que propõem e descrevem cada um dos critérios de QI (WAND; WANG, 1996; WANG; STRONG, 1996; NAUMANN; ROLKER, 2000; ZHU; BUCHMANN, 2002; CAI; ZHU, 2015). No entanto a sua aplicação é dependente daquilo que pretende-se avaliar. Em geral, são definidas métricas para avaliar os critérios de QI que geram um escore de avaliação, ou seja, um valor associado a um determinado critério de QI. As métricas são heurísticas desenvolvidas para adequar-se a uma situação de avaliação específica (PIPINO; LEE; WANG, 2002; ZAVERI et al., 2012).

Com relação aos critérios de QI, ao longo de duas décadas, muitos trabalhos propuseram suas próprias classificações. Contudo, na literatura não há um acordo sobre o conjunto de

critérios que caracterizam QI, e apesar das muitas propostas, nenhuma tem emergido como padrão. Outra observação é que não há um acordo sobre o significado dos critérios, por exemplo, é possível encontrar critérios de QI que possuem nomenclaturas diferentes mas que representam a mesma coisa, ou critérios com nomes iguais mas que representam coisas distintas.

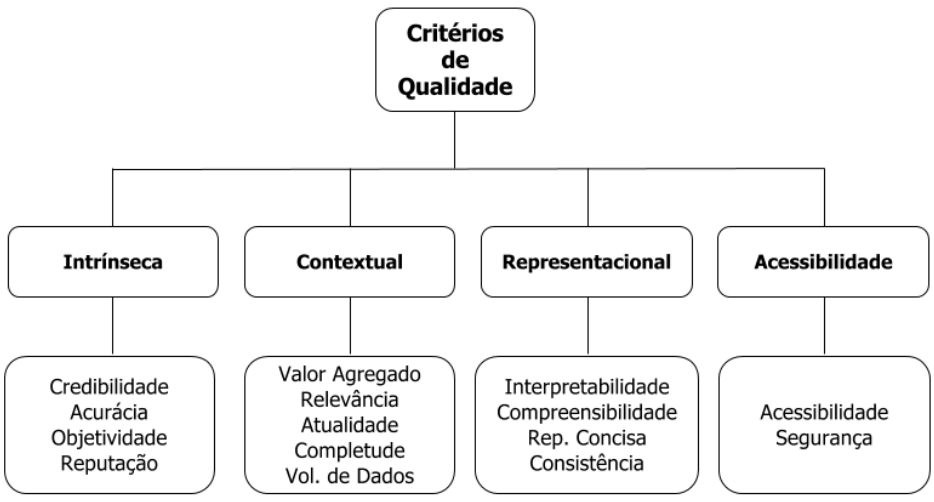
Dessa forma, fizemos um levantamento na literatura, a partir de alguns trabalhos publicados no decorrer de duas décadas, com o objetivo de apresentar as classificações dos critérios de QI ao longo do tempo, os mesmos são brevemente comentados na seção seguinte.

2.1.1 Classificação dos Critérios de QI

As primeiras pesquisas na área objetivavam a melhoria da qualidade da informação gerada pelos Sistemas de Informações (WAND; WANG, 1996; WANG; STRONG, 1996). Em Wand e Wang (1996), os autores propuseram avaliar a qualidade dos dados por meio de cinco (5) critérios: acurácia, completude, consistência, atualidade e confiabilidade. O estudo se baseou em definir métricas que mapeassem funções do mundo real para um sistema de informação. Por exemplo, diz-se que uma informação não é exata (*inaccuracy*) quando a mesma estiver representada de forma diferente do mundo real.

O trabalho de Wang e Strong (1996) se destaca por ser um dos primeiros trabalhos que oferece um conjunto de critérios de qualidade estruturados e classificados que tem sido considerado uma forte referência para a maioria dos estudos da área. Os autores identificaram, de maneira empírica, por meio de usuários que determinaram quais características dos dados eram úteis para suas tarefas, um conjunto de quinze (15) critérios, que posteriormente foram agrupados em quatro (4) categorias ou dimensões. A Figura 2.1 apresenta uma adaptação do *framework* conceitual proposto pelos autores com os critérios e suas respectivas categorias.

Figura 2.1: *Framework* conceitual - Dimensões e Critérios de Qualidade



Fonte: Adaptado de Wang e Strong (1996)

- **Intrínseca** - Os critérios relacionados à esta categoria denotam a qualidade do dado em si, independente do contexto. Por exemplo, se os dados são precisos e/ou confiáveis.

- **Contextual** - Refere-se a critérios que estão ligados a um determinado contexto, por exemplo, se são relevantes, oportunos ou completos com relação a algum requisito específico.

- **Representacional** - Esta categoria denota critérios que referem-se ao formato e significado dos dados. Por exemplo, se podem ser facilmente interpretados ou se estão representados de forma consistente.

- **Acessibilidade** - Os critérios relacionados à esta categoria definem se os dados estão disponíveis ou podem ser acessados pelo usuário. Mesmo que os dados sejam precisos e dispostos de maneira que sua interpretação seja simples, se não estiverem disponíveis possuirão pouco valor (WANG; KON; MADNICK, 1993).

A partir desses estudos, metodologias e estratégias que apoiam a melhoria da QI surgiram (BATINI et al., 2009), como é o caso da metologia *Total Data Quality Management* (TDQM) proposta em Wang (1998).

A metodologia TDQM pode ser descrita como um conjunto de teorias sólidas utilizadas para prover métodos práticos que melhorem a qualidade dos dados. O conjunto da metodologia abrange as quatro dimensões e seus respectivos critérios, apresentados em Wang e Strong (1996) e ilustrados na Figura 2.1. O processo proposto na metodologia TDQM engloba quatro (4) etapas que apoiam o ciclo da administração da qualidade dos dados (Figura 2.2), e que serão descritos a seguir.

Figura 2.2: Etapas do Processo - TDQM

Fonte: http://www.datasetting.com.br

- **Define** - Nesta etapa são definidos os requisitos de qualidade, ou seja, os critérios e normas que os dados devem seguir. Normalmente, essa etapa é definida pelo analista e por interessados em utilizar os dados. Os critérios utilizados fazem referência aos que foram categorizados nas quatro dimensões propostas por Wang e Strong (1996).

- **Mede** - A segunda etapa consiste na avaliação dos dados com base nos critérios definidos anteriormente, onde são mensurados valores para os mesmos, e/ou é possível a utilização de técnicas estatísticas para registrar dados como: taxa de erro, frequência e dispersão.

- **Analisa** - Esta etapa objetiva fazer uma análise e interpretação dos resultados obtidos, por meio da etapa anterior, e identificar possíveis ações de melhorias.

- **Melhora** - Nesta etapa, ações de melhorias da qualidade são aplicadas. Estas ações podem ser corretivas (modificações diretamente nos dados) ou preventivas (adaptando o processo que produz os dados). Em Sidi et al. (2012) essas estratégias de melhorias foram discutidas e classificadas como *data-driven* e *process-driven*, respectivamente.

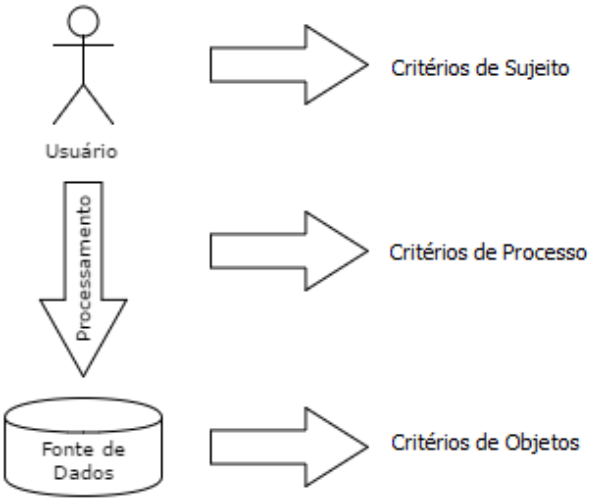
Posteriormente, com o desenvolvimento da *Web* como plataforma para publicação e consumo de dados, sistemas como os de *data warehouse* e as soluções de integração de dados se popularizaram, devido ao grande volume de dados distribuídos em diversas fontes de dados e a natureza de alta disponibilidade destes dados por parte de vários grupos de fornecedores (ARAZY; KOPAK, 2011). Nesse cenário, a avaliação da QI aplicada à fonte de dados foi alvo de pesquisas em diversos trabalhos, principalmente naqueles que aplicam a QI como meio para solucionar o problema de seleção de fontes de dados (ZHU; BUCHMANN, 2002; BASTOS et al., 2010).

Em Naumann e Rolker (2000) os autores afirmam que a avaliação da QI de uma fonte de dados deve ser realizada considerando três fatores maiores: a percepção do usuário, a fonte de dados por si só, e o acesso à fonte de dados. Esses fatores foram categorizados em três (3) classes: usuário, fonte e processamento, onde cada classe é composta por classes de conjunto de critérios, denominados de critérios de sujeito, de objetos e de processos (Figura 2.3).

- **Critérios de Sujeito** - De acordo com Naumann e Rolker (2000), o ponto de vista do usuário é um aspecto importante e que deve ser levado em consideração em uma avaliação de qualidade, uma vez que são fundamentais para decidir quais informações são boas ou não. Os critérios de sujeito normalmente são avaliados com base na experiência e/ou opinião de um ou vários usuários.

- **Critérios de Processo** - Esses critérios podem ser avaliados por meio de consultas efetuadas em uma fonte de dados e seus escores podem variar de acordo com o tipo da consulta. São critérios que podem ser avaliados automaticamente, mas que precisam ser mensurados de maneira contínua.

Figura 2.3: Classe dos Critérios de QI



Fonte: Adaptado de Naumann e Rolker (2000)

- **Critérios de Objetos** - Esses critérios tem como objetivo avaliar, por meio de uma cuidadosa análise e de forma objetiva, características relacionadas aos dados de uma fonte de dados.

O Quadro 2.1 apresenta o conjunto de critérios de QI elencados em Naumann e Rolker (2000), dentro de cada classe é especificado o método de avaliação proposto para obtenção de um valor (escore) para cada um deles.

Quadro 2.1: Classificação dos Critérios QI

Classe	Critério de QI	Método de Avaliação
Sujeito	Credibilidade	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
Objetos	Completude	Amostragem
	Objetividade	Entradas de Usuário
	Preço	Avaliação Contínua
	Confiabilidade	Contratação
	Segurança	Parsing
	Atualidade	Parsing
	Verificabilidade	Entradas de Usuário Especialista
Processo	Volume de Dados	Avaliação Contínua
	Disponibilidade	Avaliação Contínua
	Latência	Avaliação Contínua
	Tempo de Resposta	Avaliação Contínua

Fonte: Adaptado de Naumann e Rolker (2000)

Em síntese, os métodos de avaliação dos critérios de QI são:

- **Experiência do Usuário** - Pode ser mensurado por meio do conhecimento do usuário

adquirido ao longo do tempo com relação às fontes de dados, por exemplo: uso de relatórios e sistemas de informações, ou a própria experiência a partir de seu uso.

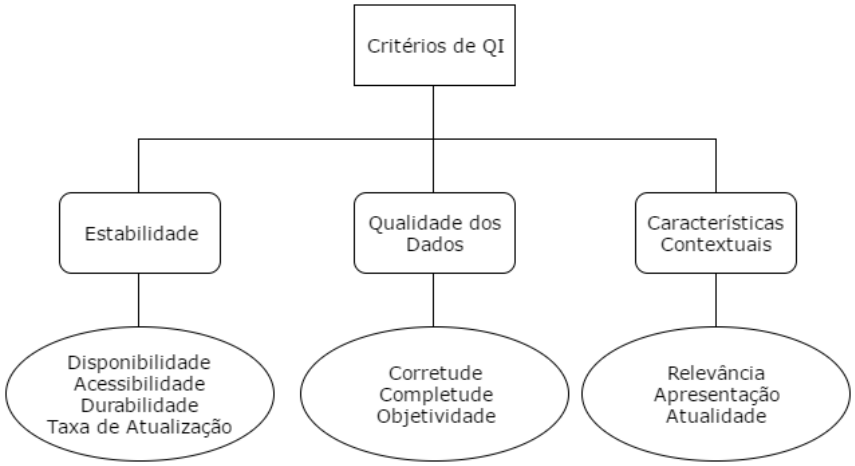
- **Amostragem do Usuário** - Com esse método o usuário utiliza uma amostra de resultados obtidos das fontes de dados para realizar uma avaliação do critério.
- **Avaliação Contínua do Usuário** - Trata-se de fazer uma verificação contínua das informações que o usuário recebe da fonte de dados ao longo do tempo. Nesse sentido, o usuário analisa cada informação recebida, não só uma amostragem.
- **Parsing** - Efetuar operações de *parsing* significa extrair e processar metadados das fontes de dados. Os autores consideram que as fontes de dados fornecem documentações referentes ao seu conteúdo.
- **Amostragem** - A fim de evitar a tarefa de avaliar todo o conteúdo de uma fonte de dados na intenção de estimar um escore correto de um critério, técnicas de amostragem são aplicadas para selecionar um subconjunto representativo das informações e considerar apenas esse subconjunto na determinação do escore do critério de qualidade.
- **Entrada de Usuário Especialista** - O conhecimento de usuários especialistas é necessário para estimar alguns critérios, por exemplo, o critério de verificabilidade (alguns autores chamam de Corretude). Nesse sentido, o especialista deve seguir algumas diretrizes para garantir a precisão e comparação dos escores.
- **Avaliação Contínua** - Trata-se de uma avaliação periódica realizada nas fontes de dados com o objetivo de verificar quão bem as mesmas estão com relação a cada critério.

O trabalho de Zhu e Buchmann (2002) também propõe uma classificação dos critérios de QI para avaliação de fontes de dados. A classificação é apresentada na Figura 2.4.

As categorias foram definidas como segue:

- **Estabilidade** - Essa categoria engloba critérios cujo objetivo é medir a qualidade do serviço provido pela fonte de dados. Exemplos de critérios de QI nessa categoria são: disponibilidade e taxa de atualização.
- **Qualidade dos Dados** - Essa categoria engloba critérios cujo objetivo é medir a qualidade dos dados contidos na fonte de dados, tais como: corretude e completude.
- **Características Contextuais** - Essa categoria é formada por critérios mais subjetivos, cujo objetivo é medir a qualidade da fonte de dados com base no contexto da sua utilização. Exemplos de critérios nessa categoria são: relevância e atualidade.

Figura 2.4: Classificação dos Critérios de QI

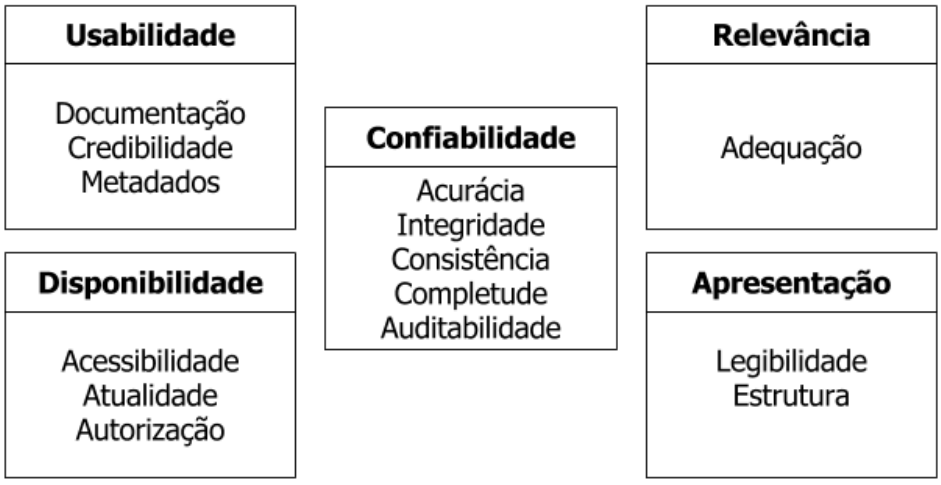


Fonte: Adaptado de Zhu e Buchmann (2002)

Em Cai e Zhu (2015), os autores trouxeram à tona uma realidade mais próxima dos dias atuais, o *big data*. A atual capacidade de gerar e coletar dados resulta em um grande volume de dados que precisam ser tratados e disponibilizados para gerar informação. O valor da informação está relacionado diretamente com a sua qualidade. Nesse cenário, o referido trabalho além de fazer um levantamento sobre os desafios de avaliar qualidade na era *big data*, ainda propõe um *framework conceitual* com um conjunto de critérios de QI, cujo objetivo é definir um processo de avaliação dinâmica da qualidade dos dados.

O *framework* é composto por um conjunto de dimensões (categorias para agrupar os critérios), elementos (critérios de QI) e indicadores (métricas de avaliação dos critérios de QI). As dimensões e critérios de QI foram selecionados com base nos 5V do *big data*: Volume, Velocidade, Veracidade, Valor e Variedade. A Figura 2.5 apresenta as dimensões e elementos definidos pelos autores.

Figura 2.5: Dimensões e Elementos de QI



Fonte: Adaptado de Cai e Zhu (2015)

2.1.2 Critérios de QI Relevantes para Avaliação de uma Fonte de Dados

Nesta seção, são apresentados alguns critérios e métricas de QI que podem ser utilizados para avaliar a qualidade de uma fonte de dados, de acordo com a visão dos principais autores da área.

- **Disponibilidade** - Esse critério é relevante principalmente no contexto da *Web*. Normalmente, as fontes de dados na *Web* possuem um serviço que disponibilizam seus dados. Nesse contexto, não há um controle sobre as fontes de dados, ou seja, não há garantias de que elas manterão sempre seus serviços disponíveis. Naumann (1998) propõe como método para avaliar esse critério a submissão de consultas que atestem a disponibilidade da fonte de dados.

  Nesse caso, fontes de dados com um alto escore para o critério disponibilidade tendem ser melhores com relação à utilização dos seus dados para consumo. Naumann e Rolker (2000) dizem que o escore para disponibilidade de uma fonte de dados pode ser dado como sendo o percentual de tempo no qual uma fonte de dados permanece disponível.

- **Tempo de Resposta** - De acordo com Zaveri et al. (2012), o critério tempo de resposta equivale à duração entre a solicitação de um pedido por parte do usuário e a recepção da resposta de um sistema. Dessa forma, para avaliar esse critério é necessário submeter uma consulta à fonte de dados e recuperar o tempo que foi gasto para a mesma responder tal consulta. Batista (2008) sugere que esse tipo de avaliação seja feita de forma automática por meio de estatísticas sobre o tempo médio de resposta obtido em diferentes momentos.

- **Atualidade** - O critério atualidade possui múltiplas vertentes. Muitos autores se referem ao mesmo critério utilizando objetivos diferentes. Por exemplo, em Zhu e Buchmann (2002) esse critério está relacionado com a frequência com que a fonte de dados se atualiza, em outras palavras, verifica-se a "pontualidade"com que a fonte de dados é atualizada. Em Wang e Strong (1996), esse critério avalia o dado propriamente dito. Nesse sentido os autores definem o critério como “o grau em que a idade dos dados é apropriada para a tarefa em mãos”, ou seja, se os dados são atuais para o que se deseja fazer. Em Bouzeghoub e Peralta (2004) os autores propõem analisar a atualidade do dado com base em duas perspectiva: *currency*, ou seja, o tempo entre a extração do dado da fonte e a entrega para o usuário, e *timeliness*, que em outras palavras verifica se o dado está atual para o uso. Nesse sentido, foram elencadas métricas específicas para cada tipo de perspectiva.

- **Reputação** - Wang e Strong (1996) definem esse critério como “a medida que os dados são confiáveis ou altamente consideráveis em termo de origem e conteúdo”.

Na verdade, esse é um critério totalmente subjetivo e baseado em avaliações que os usuários fazem. Podemos dizer que uma fonte de dados possui uma boa reputação quando a mesma possui dados confiáveis, credíveis e livres de erro. Naumann e Rolker (2000) sugerem que a avaliação para esse critério seja feita considerando uma pontuação atribuída pelos usuários por meio de uma nota entre 1 (má reputação) e 10 (boa reputação).

- **Corretude** - Zhu e Buchmann (2002) definem o critério corretude como o grau em que os dados estão livres de erros, já Wang e Strong (1996) utilizam o termo *verificabilidade* para identificar a medida que os dados podem ser verificados quanto à sua correção. Para isso, os autores propõem que esse critério seja medido por especialistas do domínio. Em trabalhos como o de Xian et al. (2009) esse critério é mensurado de forma objetiva, por exemplo, um conjunto de restrições são definidas e aplicadas aos dados. O grau de corretude pode ser encontrado pelo total de dados que estão corretos dividido pelo total de dados avaliados.

- **Completude** - O conceito de completude pode ser dado como "o grau em que os dados são suficientes" para uma determinada tarefa (WANG; STRONG, 1996). No entanto, em Pipino, Lee e Wang (2002) esse conceito se expandiu e foi proposto avaliar a completude dos dados e do esquema das fontes. Com relação aos dados, diz-se que uma fonte de dados é completa se ela não possui dados nulos ou faltantes. Já com relação ao esquema, diz-se que uma fonte de dados é completa se possui em seu esquema todos os atributos da aplicação.

  A métrica para medir a Completude de Dados consiste na divisão da quantidade de dados incompletos pela quantidade total de dados, cujo resultado é subtraído de um (1) (XIAN et al., 2009). A métrica para medir a Completude do Esquema pode ser dada pela quantidade de itens contidos no esquema da fonte que estejam presentes no esquema da aplicação, dividido pelo total de itens contidos no esquema da aplicação (LOSCIO et al., 2012).

- **Consistência** - O critério consistência também pode ser visto por dois prismas: em termos de dados e de representação. Diz-se que um dado é consistente se não apresenta nenhum tipo de conflito. Por exemplo, Xian et al. (2009) consideram que uma fonte de dados é consistente se não possui dados redundantes. Logo, o valor para o critério é a razão entre o número de itens inconsistentes e o número total de itens, subtraindo de um.

  Quanto à questão da representação dos dados, em Wang e Strong (1996) é proposto o critério Consistência Representacional, que pode ser definido como a medida que os dados são sempre apresentados em um mesmo formato. Uma forma de avaliar

esse critério é verificando se os dados contidos na fonte se apresentam em um mesmo formato tal qual definido em seu esquema.

2.2 Seleção de Fontes

O desenvolvimento da *Web* como plataforma para publicação e consumo de dados possibilitou um aumento no número de fontes de dados disponíveis para os usuários. Esta alta disponibilidade trouxe à tona um grande problema: como encontrar as melhores fontes de dados para uma necessidade específica? Nesse sentido, o problema de seleção de fontes de dados vem sendo bastante discutido nos últimos anos e amplamente abordado em algumas áreas da Ciência da Computação, como Recuperação da Informação (RI) e Integração de Dados.

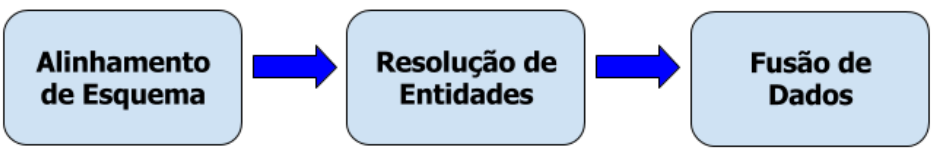
Em RI o objetivo é consultar apenas as fontes de dados mais adequadas às necessidades do usuário ou de uma determinada aplicação. Nesse contexto, normalmente são utilizadas soluções baseada em *metasearch*.

O *metasearch engine* é um mecanismo de busca que envia uma consulta para múltiplas fontes e reúne os resultados em um único conjunto que é apresentado para o usuário que realizou a consulta. No entanto, um componente importante nesse tipo de solução é o de seleção de fontes. O objetivo desse componente é selecionar um conjunto de fontes mais prováveis de fornecer uma informação relevante para o usuário, diminuindo assim o escopo de fontes de dados que serão usadas nas consultas.

Em Aksoy (2005) são discutidas algumas abordagens que podem ser implementadas em soluções baseada em *metasearch*, tais como: busca por palavras-chaves e consultas. Em Samir e Habiba (2009), os autores propõem a criação de um perfil da fonte de dados, com algumas características tais como: Formato, Linguagem, entre outras e um perfil do usuário que está fazendo a busca, com informações de suas preferências. A partir disso, aplica-se técnicas de similaridade que compara os dois perfis buscando encontrar as fontes mais adequadas para cada tipo de consulta de acordo com o objetivo e características do usuário.

No contexto de Integração de Dados a seleção de fontes é vista como uma alternativa para melhorar o processo de integração. Dong e Srivastava (2015) destacam três principais etapas para o processo de Integração de Dados, são elas: Alinhamento de Esquema (*Schema Alignment*), Resolução de Entidades (*Record Linkage*) e Fusão de Dados (*Data Fusion*) (Figura 2.6).

Figura 2.6: Etapas do Processo de Integração de Dados



Fonte: Adaptado de Dong e Srivastava (2015)

O processo de Alinhamento de Esquema consiste em realizar a correspondências de

esquema entre fontes de dados distintas (DONG; SRIVASTAVA, 2015). Resolução de Entidades busca identificar diferentes referências para uma mesma entidade no mundo real (BHARAMBE; JAIN; JAIN, 2012). Fusão de Dados busca resolver esses conflitos de referências das entidades, fornecendo um valor integrado para cada instância da entidade (DONG; NAUMANN, 2009).

O interesse no desenvolvimento de aplicações que utilizam dados integrados de múltiplas fontes tem experimentado um crescimento enorme nos últimos anos, principalmente por conta do grande número de fontes de dados disponíveis. No entanto, integrar um grande número de fontes de dados torna-se uma tarefa dispendiosa. As fontes podem possuir variação em relação à sua qualidade e o custo para recuperar e resolver problemas de inconsistências de dados, mediação de esquema, além de tratar outros aspectos relacionados ao processamento desses dados pode ser alto (REKATSINAS et al., 2015).

Uma vez que se tem um grande número de fontes de dados, os custos para acessar a fonte, processar os dados, executar consultas, realizar o mapeamento dos esquemas, entre outras atividades tendem a ser altos. Nesse contexto, algumas estratégias têm sido propostas pra identificar as melhores fontes para serem integradas, como o uso do *feedback* (TALUKDAR; IVES; PEREIRA, 2010; OLIVEIRA, 2012) e critérios de QI para avaliar as fontes de dados (XIAN et al., 2009; DONG; SAHA; SRIVASTAVA, 2012; DUSTDAR et al., 2012; REKATSINAS; DONG; SRIVASTAVA, 2014; REKATSINAS et al., 2015).

2.2.1 Uso de Critérios de QI para Seleção de Fontes de Dados

A utilização de critérios de QI no processo de seleção de fontes de dados consiste em avaliar, por meio de um conjunto de critérios de QI, um conjunto de fontes de dados candidatas ao uso. Uma avaliação de qualidade deve ser multidimensional, como o próprio conceito define, ou seja, deve-se considerar múltiplos critérios de QI (NAUMANN; FREYTAG; SPILIOPOULOU, 1998; BATISTA, 2008). Por exemplo, se a avaliação considera apenas o critério disponibilidade, alguma fonte pode obter um excelente escore para este critério, no entanto, a mesma pode não ser tão boa quanto à corretude de seus dados.

Na Seção 2.1.2, apresentamos alguns critérios de QI que podem ser utilizados para avaliar a qualidade de uma fonte de dados. Uma vez que os critérios de QI foram selecionados, cada um deve está associado a uma métrica, para que seja possível mensurar um valor ou escore para eles. Entretanto, quando estamos no contexto de seleção de fontes de dados, além de avaliar é necessário classificar o conjunto de fontes de dados avaliados.

Normalmente, a classificação é realizada mediante a combinação dos valores atribuídos aos critérios que foram avaliados (XIAN et al., 2009; SARINHO, 2014). Entretanto, comparar múltiplos critérios torna-se um problema de decisão de multicritérios (ZHU; BUCHMANN, 2002).

Múltiplos critérios geralmente têm unidades de medidas diferentes. Por exemplo, o critério tempo de resposta pode ser mensurado em *milissegundos*, enquanto a reputação pode

ser representada com valores entre 0 a 10 (Seção 2.1.2). Alguns critérios de QI também podem ser classificados como positivo/qualidade (quanto maior o valor para esse critério melhor a fonte é qualificada) ou negativo/custo (quanto menor o valor para esse critério melhor a fonte é qualificada) (NAUMANN; FREYTAG; SPILIOPOULOU, 1998). Utilizando o exemplo acima, o critério reputação seria classificado como positivo e tempo de resposta como negativo.

Dessa forma, torna-se necessário uniformizar os valores atribuídos aos critérios de QI a fim de torná-los comparáveis. Em Naumann, Freytag e Spiliopoulou (1998) e Zhu e Buchmann (2002) são apresentados métodos que solucionam esse problema. Na próxima seção detalharemos os métodos propostos.

2.2.2 Métodos para Tomada de Decisão de Múltiplos Critérios

Os métodos para tomada de decisão de múltiplos critérios podem ser classificados em quatro (4): (i) Poderação Aditiva Simples; (ii) Técnica para ordenar preferências por similaridade com a solução ideal; (iii) Processo de análise hierárquica; e (iv) Análise por envoltória de dados (NAUMANN; FREYTAG; SPILIOPOULOU, 1998; ZHU; BUCHMANN, 2002).

- **Poderação aditiva simples** - *Simple Additive Weighting* (SAW) - É um dos métodos mais populares e bem aceitos. Apesar de ser um método mais simples, os resultados alcançados por meio de sua utilização são bem próximos de outros obtidos com métodos mais sofisticados. O método é dividido em três etapas: (i) normalização dos escores obtidos em cada critério para torná-los comparáveis; (ii) aplicar peso a cada um dos critérios; e (iii) somar os valores dos escores de cada fonte de dados. Em Zhu e Buchmann (2002), o cálculo para normalização e combinação dos escores são apresentados.

- **Técnica para ordenar preferências por similaridade com a solução ideal** - *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS) - É uma abordagem voltada para encontrar uma alternativa que seja a mais próxima da solução ideal e mais afastada da solução ideal-negativa em um espaço de computação multidimensional. Esse espaço é determinado pelo conjunto de critérios de qualidade. A solução ideal representa o conjunto dos melhores escores dos critérios e a solução ideal-negativa com os piores escores (ZHU; BUCHMANN, 2002).

- **Processo de análise hirárquica** - *Analytical Hierarchy Process* (AHP) - É composto por quatro passos: (i) desenvolvimento de uma hierarquia de objetivos. Neste contexto, o objetivo principal é a seleção de fontes. A hierarquia é construída a partir dos critérios de qualidade; (ii) comparação em pares de objetivos, que é feita para avaliar a importância relativa entre os objetivos; (iii) verificação da consistência das comparações, no qual é gerada uma matriz de comparações entre os pares de objetivos (critérios de qualidade); e (iv) agregação das comparações obtidas, em que

para cada critério é feita a normalização. Por fim, calcula-se a média ponderada, gerando uma medida global de qualidade usada para classificar as fontes de dados (ZHU; BUCHMANN, 2002).

- **Análise por envoltória de dados** - *Data Envelopment Analysis* (DEA) - Esse método não fornece uma classificação das fontes de dados. Ele é utilizado para encontrar uma ou mais fontes de dados com o maior escore dentro do conjunto de critérios de qualidade. Assim, os pesos não são especificados pelos usuários, eles são variáveis determinadas na execução do método (NAUMANN; FREYTAG; SPILIOPOULOU, 1998; ZHU; BUCHMANN, 2002).

Em Sarinho (2014) é feito um levantamento sobre esses diferentes métodos, onde além de mostrar a sua utilização é feita uma comparação. O estudo apontou que quando a avaliação é realizada utilizando uma quantidade pequena de critérios de qualidade os métodos SAW e TOPSIS são mais indicados por serem mais práticos e terem um custo de processamento menor no momento de classificar as fontes, diferentemente do AHP, que precisa gerar uma matriz de comparação para classificar as fontes. O DEA é indicado quando a situação for contrária. Sua grande vantagem é que o processo de seleção é feito de forma mais objetiva, sem a indicação de pesos para cada critério. No entanto, a grande desvantagem é que não é fornecida uma classificação das fontes de dados.

2.3 *Data Profiling*

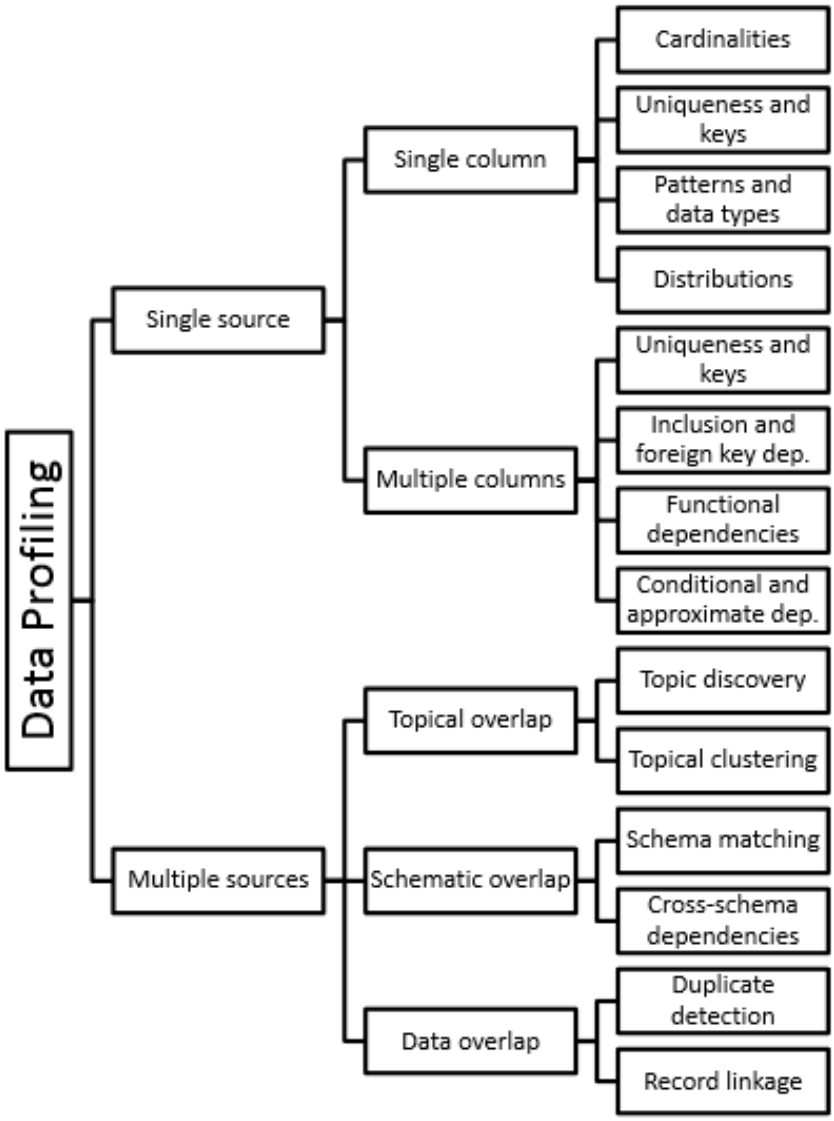
Data Profiling é um conjunto de tarefas e processos para analisar e produzir metadados sobre um determinado conjunto de dados (ABEDJAN; GOLAB; NAUMANN, 2015). Johnson (2009) define como a atividade de criar informações sumarizadas de uma base de dados.

Essas informações, geradas em forma de metadados utilizando tarefas de *Data profiling*, podem ser utilizadas para vários fins. Dentre elas, pode-se citar:

- Permitir que usuários entendam melhor a estrutura, o escopo e a distribuição dos valores presente em uma fonte de dados;
- Auxiliar no processo de *data cleansing*, por meio da identificação de valores errôneos, faltantes e *outlier* (WANG; DONG; MELIOU, 2015);
- Auxiliar na tarefa de análise de dados. Muitas análises envolvem o uso de estatísticas. Os metadados coletados podem servir de apoio para tais análises.

Em Naumann (2014) é realizada uma classificação das tarefas de *data profiling*. As tarefas são classificadas em: única fonte (*single source*) e múltiplas fontes (*multiple sources*). As tarefas para única fonte correspondem a tarefas tradicionais encontradas na literatura, já as tarefas

Figura 2.7: Classificação das tarefas de *Data profiling*



Fonte: Naumann (2014)

para múltiplas fontes correspondem a novas direções de pesquisas para a área, apresentadas como contribuições do trabalho citado (Figura 2.7).

As tarefas para uma única fonte de dados podem ser divididas em duas: (i) única coluna (*Single column*); e (ii) múltiplas colunas (*Multiple columns*). As tarefas de única coluna envolvem metadados mais simples. Entre eles destacam-se os que são obtidos por meio de métodos estatísticos, tais como: máximo, mínimo, desvio padrão, frequência de distribuição, entre outros. Esses métodos são utilizados para encontrar, por exemplo, número de valores nulos e distintos de uma coluna, tipos de dados e padrões de valores mais frequentes em uma coluna.

Outras verificações mais complexas, utilizando múltiplas colunas de um conjunto de dados, também podem ser feitas, como é o caso de analisar a correlação de dependência entre as colunas e a identificação de correlações de valores por meio de regras de asssociação (ABEDJAN;

GOLAB; NAUMANN, 2015). Normalmente, esses metadados são úteis em aplicações que fazem exploração e análise de dados como meio para detectar *outliers*.

As tarefas para múltiplas fontes de dados são divididas em três: (i) sobreposição de tópicos (*Topical overlap*); (ii) sobreposição de esquemas (*Schematic overlap*); e (iii) sobreposição de dados (*Data overlap*). Tais tarefas permitem que, no contexto de Integração de Dados, sejam gerados metadados que possibilitem saber sobre a sua "integrabilidade", isto é, quão bem os dados e os esquemas das múltiplas fontes de dados analisadas se encaixam.

Em Naumann (2014) e Abedjan, Golab e Naumann (2015), são elencados diversos métodos que podem ser utilizados para as tarefas de *data profiling* citadas. É importante destacar que existem várias ferramentas que surgiram por meio de pesquisas na área, como é o caso da Bellman (DASU et al., 2002), Potters Wheel (RAMAN; HELLERSTEIN, 2001), MADLib (HELLERSTEIN et al., 2012), entre outras. Ferramentas comerciais também podem ser encontradas, como é o caso do SQL Server Data Profiling Task, comercializada pela Microsoft[1] e Enterprise Data Quality, comercializada pela Oracle[2].

2.4 Considerações

Neste capítulo apresentamos os principais tópicos relacionados a esta dissertação. Primeiramente, abordamos o conceito de QI seguindo pela apresentação de alguns trabalhos encontrados na literatura que propõem e classificam critérios de QI. Em seguida, destacamos alguns critérios de QI relevantes para avaliação de uma fonte de dados, fazendo uma breve sumarização da definição e métricas para avaliar cada um deles sob a ótica dos principais autores da área.

Posteriormente, discutimos o problema de seleção de fontes de dados, onde enfatizamos o uso de critérios de QI como método para tal problema. Por fim, introduzimos o conceito de *data profiling*. Nesta pesquisa, nos inspiramos em tal conceito para a definir o Perfil de Qualidade que estamos propondo. Nesse sentido, as principais tarefas e aplicações de uso de *data profiling* foram destacados. O próximo capítulo apresenta os trabalhos que se relacionam com o contexto desta dissertação.

[1] https://www.microsoft.com
[2] https://www.oracle.com/

3

Trabalhos Relacionados

Este capítulo tem como objetivo apresentar e discutir alguns trabalhos existentes na literatura que possuem relação com a proposta desta pesquisa. Dividimos os trabalhos em duas categorias, a primeira elenca trabalhos com foco em avaliar a qualidade das fontes de dados e a segunda elenca trabalhos que propõem abordagens para geração e utilização de metadados das fontes de dados para diversos fins. Os trabalhos selecionados serão descritos nas seções seguintes, e uma breve discussão será realizada em seguida.

3.1 Avaliação da Qualidade em Fontes de Dados

A literatura oferece diversos trabalhos cujo foco é avaliar a qualidade das fontes de dados para os mais diversos fins. Selecionamos alguns desses trabalhos e apresentamos uma sumarização de cada um.

3.1.1 Qualidade de Esquema em um Sistema de Integração de Dados

O principal objetivo do trabalho proposto por Batista (2008) é melhorar a qualidade da execução das consultas submetidas a um sistema de integração de dados. Essa hipótese se baseia no fato que uma alternativa para otimizar o processamento das consultas seria a construção de esquemas com bons escores de qualidade. Para isso, o trabalho propõe analisar a qualidade dos esquemas de integração de dados, mais especificamente o esquema integrado.

Geralmente o esquema integrado é gerado por módulos internos do sistema de integração de dados. No contexto do referido trabalho, o esquema integrado é formado por um conjunto de entidades, na qual cada entidade pode possuir um conjunto de atributos, e também pelos relacionamentos entre as entidades. Portanto, logo após ser gerado, o esquema integrado é avaliado com base em três critérios: **minimalidade, completude de esquema e consistência de tipo**.

Minimalidade é definido como o grau de ausência de elementos redundantes no esquema. Um esquema é considerado mínimo se todos os conceitos são descritos apenas uma única vez. Para calcular o grau de minimalidade é necessário identificar o grau de entidades redundantes

bem como o grau de relacionamentos redundantes no esquema. Uma vez que esses valores são calculados eles são somados e o resultado é subtraído de 1.

O trabalho ainda propõe um algoritmo cujo objetivo é melhorar o escore da qualidade do esquema integrado para o critério de minimalidade. É apresentada como solução a eliminação de entidades, atributos e relacionamentos redundantes, uma vez que são esses os fatores para baixa pontuação em tal critério.

Completude de Esquema é descrito como o grau em que os conceitos apresentados no esquema integrado estão representados em todos os esquemas das fontes de dados do sistema de integração. O cálculo para o critério de completude de esquema consiste na quantidade de conceitos extraídos do esquema integrado dividido pelo total de conceitos encontrados em todas as fontes de dados do sistema de integração.

Por fim, o critério consistência de tipo é avaliado. Um esquema integrado é consistente se todos os atributos equivalentes são representados com o mesmo tipo de dados. Sendo assim, foi atribuído um padrão de tipo de dados para cada atributo. A métrica para o critério em questão verifica se todos os atributos equivalentes seguem o mesmo padrão. Se existir algum tipo de inconsistência o atributo recebe o valor 0, caso contrário 1. O cálculo para o critério consiste na soma dos valores atribuídos a cada atributo divido pela quantidade de atributos do esquema integrado.

A validação do trabalho foi realizada por meio de alguns experimentos, dentre eles podemos destacar um experimento que seguiu as seguintes etapas: (i) inicialmente as consultas foram submetidas a um esquema integrado com redundância; (ii) depois o algoritmo para eliminação de redundância foi executado sobre o esquema integrado, gerando um alto escore para o critério de minimalidade; (iii) as mesmas consultas foram submetidas no esquema ajustado; e (iv) foram realizadas as comparações entre os tempos gastos nas consultas submetidas na etapa (i) e (iii).

Por fim, a autora pôde concluir sua hipótese inicial: um esquema integrado com um bom escore de qualidade tende a ser mais eficiente durante o processamento de consultas em um Sistema de Integração.

3.1.2 Seleção de Fontes de Dados Baseado em Qualidade para Integração de Dados na *Deep Web*

Motivados pelo crescente número de informações úteis escondidas por trás das páginas na *Web*, o trabalho de Xian et al. (2009) propõe uma abordagem para selecionar fontes de dados com base na sua qualidade. O contexto relacionado é o de Integração de Dados.

Os autores argumentam que ao ser definido um sistema de Integração de Dados, uma das principais preocupações que se deve ter é quais fontes de dados devem ser incluídas. O escopo da *Deep Web* é bem abrangente, ou seja, é possível encontrar diversas fontes disponíveis que podem prover dados para um mesmo domínio.

A discusão do trabalho gira em torno de que a utilização de todas as fontes de dados disponíveis pode trazer resultados ineficientes, uma vez que podem existir fontes de dados com baixa qualidade. Além disso, outros custos associados, como, o de processamento e de rede para a recuperação dos dados das fontes, estão atrelados como justificativa para realizar o processo de seleção utilizando a abordagem proposta.

Sendo assim, o principal problema abordado pelos autores é o de como selecionar um conjunto de fontes de dados para um sistema de integração. Para auxiliar a decidir quais fontes de dados poderão ser incluídas em tal sistema, foi proposta uma avaliação nas fontes de dados utilizando alguns critérios de qualidade.

Primeiramente, por meio de uma análise para caracterizar as fontes de dados, foram identificados quatro critérios de qualidade como sendo relevantes para o contexto do trabalho. Depois foi proposto um modelo de avaliação utilizando métricas para cada um. Os critérios selecionados foram: **completude e consistência dos dados, tamanho da fonte de dados (volume de dados) e tempo de resposta**.

A completude dos dados é definida como o grau em que os valores das instâncias em uma fonte de dados estão presentes. O cálculo para esse critério consiste primeiramente em encontrar a porcentagem de valores faltantes e em seguida subtrair o resultado de 1 para fazer a conversão e encontrar o grau de completude da fonte de dados.

A consistência dos dados é definida como o grau em que os valores das instâncias estão livres de conflitos. Para tanto, existem regras de restrições que realizam esse tipo de verificação a fim de obter um valor para esse critério. O cálculo para esse critério consiste primeiramente em encontrar a porcentagem de valores inconsistentes e em seguida subtrair o resultado de 1 para fazer a conversão e encontrar o grau de consistência da fonte de dados.

No texto dos autores não ficou claro como as regras de restrições para o critério Consistência dos Dados foram implementadas, o que nos leva a crer que essas regras são dependentes do domínio ao qual as fontes de dados pertencem.

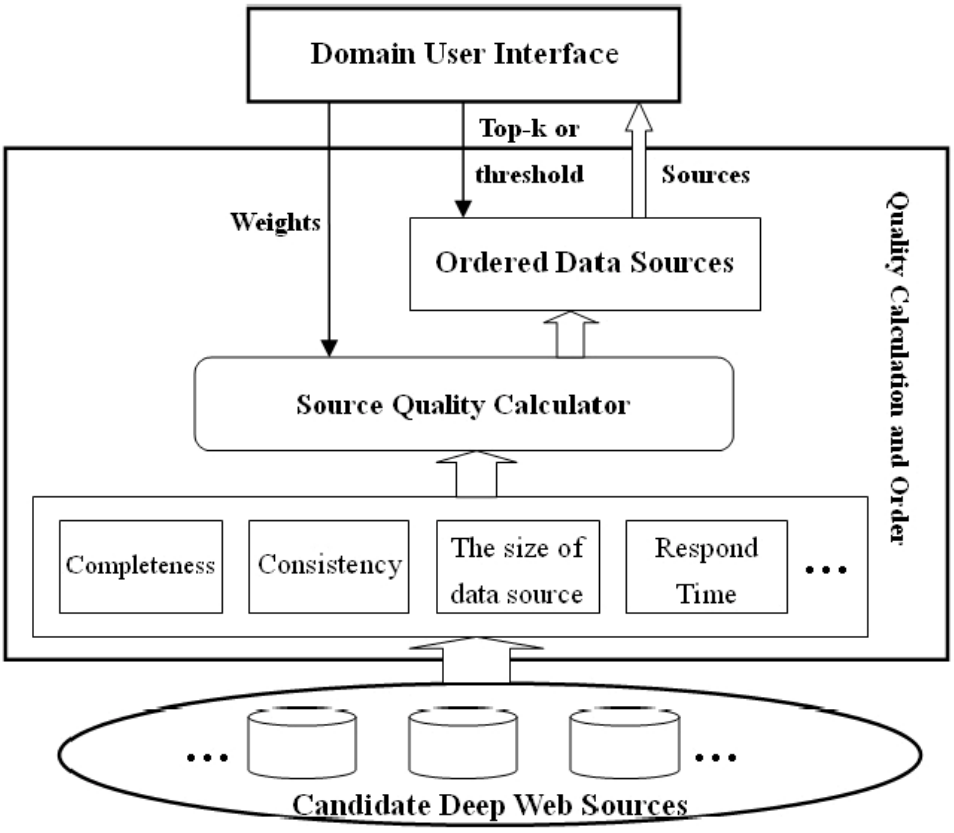
Outro critério avaliado foi o tamanho da fonte de dados. Segundo os autores, uma fonte de dados que possui maior tamanho tende ser mais bem qualificada do que uma que apresenta um tamanho inferior, uma vez que eles acreditam que o volume de dados influencia na quantidade de informações utéis que a fonte pode oferecer. Para calcular um valor para esse critério é necessário conhecer o tamanho de todas as fontes de dados que estão sendo avaliadas. O cálculo para esse critério é o tamanho da fonte de dados em questão dividido pelo maior tamanho de todas as n fontes de dados avaliadas.

O tempo de resposta também foi um dos critérios avaliados nesse trabalho. Para calcular um valor para esse critério, primeiramente foi necessário achar o tempo de resposta médio para cada fonte de dados avaliada. Para isso, foi submetido um conjunto de consultas $RQ = \{q_1, q_2,...q_n\}$ nas fontes de dados. O tempo médio de resposta para uma fonte de dados foi calculado baseado na soma de todos os tempos gastos para executar cada $q_n \in RQ$ dividido pela quantidade de consultas submetidas.

De posse de todos os tempos médios das fontes de dados avaliadas, o cálculo para esse critério consiste em dividir o tempo médio da fonte de dados em questão pelo maior tempo médio de todas as fontes de dados avaliadas, cujo resultado é subtraído de 1.

O modelo de seleção proposto pelos autores é baseado na classificação das fontes de dados (Figura 3.1). Por exemplo, dado que temos um conjunto de fontes de dados $S = \{s_1, ..., s_n\}$, o objetivo é selecionar de S apenas as fontes de dados que são *top-k* de maior qualidade ou que tenham um escore de qualidade maior ou igual a um *threshold* que foi definido.

Figura 3.1: Modelo de Seleção de Fontes de Dados



Fonte: Xian et al. (2009)

O escore de qualidade de cada fonte de dados é calculado com base nos valores obtidos para cada critério durante a avaliação da fonte de dados. Também é necessário que o usuário atribua pesos aos critérios de avaliação. O peso atribuído a um critério corresponde ao valor de importância que ele exerce na avaliação. A soma dos pesos não pode ser maior que 1. Dessa forma, o escore é calculado com base na soma de todos os valores dos critérios de qualidade que foram avaliados anteriormente multiplicados pelo peso atribuído a cada um. O escore de qualidade é um valor que varia entre 0 e 1, onde 1 representa o melhor índice de qualidade e 0 o pior.

3.1.3 Usando Informação de Qualidade para Identificar Fontes de Dados Web Relevantes: Uma Proposta

O trabalho de Loscio et al. (2012) discute o problema da identificação de fontes de dados relevantes para aplicações de domínio específico. Para auxiliar na identificação, as autoras propõem o uso de informações de qualidade das fontes de dados.

Primeiramente, para identificar o grau de relevância de uma fonte de dados com relação a uma aplicação, é necessário: (i) descrever tanto as fontes de dados como os requisitos da aplicação e (ii) definir quais critérios de QI que podem ser utilizados a fim de ajudar no cálculo de relevância. Especificamente, o enfoque principal do trabalho é a segunda questão levantada.

No entanto, com relação à primeira questão, a respeito da descrição das fontes e dos requisitos da aplicação, as autoras assumem a utilização de ontologias de domínio para descrever os conceitos presentes nas fontes de dados.

Os requisitos da aplicação são descritos por um conjunto de consultas $Q = \{q_1, ..., q_n\}$, onde cada consulta $q_n \in Q$ possui um conjunto de conceitos que as descreve. Como falamos anteriormente, cada fonte de dados também possui um conjunto de conceitos que são responsáveis pela sua descrição. Sendo assim, dado que temos um conjunto $S = \{s_1, ..., s_m\}$ de fontes de dados candidatas à aplicação, uma avaliação é realizada para descobrir o grau de relevância baseado em três critérios de QI: **completude de esquema, completude e corretude dos dados**.

A completude de esquema diz respeito à quantidade de informações que é possível obter de uma fonte de dados com relação aos requisitos da aplicação. Para calcular o valor para esse critério é levado em consideração a quantidade de conceitos encontrados na fonte de dados com relação aos conceitos que estão presentes nas consultas. Como maneira de enriquecer esse cálculo, as autoras incorporaram um peso para cada conceito encontrado nas fontes de dados. Esse peso é definido de acordo com o grau de similaridade entre os conceitos das consultas e os conceitos da fonte de dados. O grau de similaridade é calculado utilizando a abordagem *SemMatcher* (PIRES et al., 2009).

A completude de dados nesse trabalho é definida como a razão entre o tamanho do conjunto de resposta da consulta em relação à quantidade total de dados conhecidos. Em outras palavras, a métrica utilizada para calcular o valor desse critério é a quantidade de instâncias retornadas pela fonte de dados que está sendo avaliada dividido pelo total de instâncias que foi retornada por todo o conjunto de fontes de dados S avaliado.

Por fim, o último critério avaliado é a corretude dos Dados. Ele é definido como o grau em que os dados estão livres de erro. É necessário saber a quantidade de erros totais encontrados em todas as fontes de dados que estão sendo avaliadas. A métrica utilizada é a divisão da quantidade de erros encontrados na fonte de dados em questão sobre o total de erros de todo o conjunto S de fontes de dados subtraído de 1.

Uma vez que os critérios foram calculados para todas as fontes de dados é gerada uma medida de relevância. Essa medida tem como objetivo gerar um valor entre 0 e 1 e será utilizada

para classificar as fontes de dados contidas em S. A medida de relevância é a informação que será usada pelo desenvolvedor para identificar as melhores fontes de dados para sua aplicação.

O cálculo de relevância de uma fonte de dados com relação a uma aplicação é constituído pela multiplicação dos valores obtidos para os critérios completude de esquema (CE) e completude de dados (CD), somado pelo valor do critério de corretude (C) cujo resultado é dividido por dois, conforme a Equação 3.1.

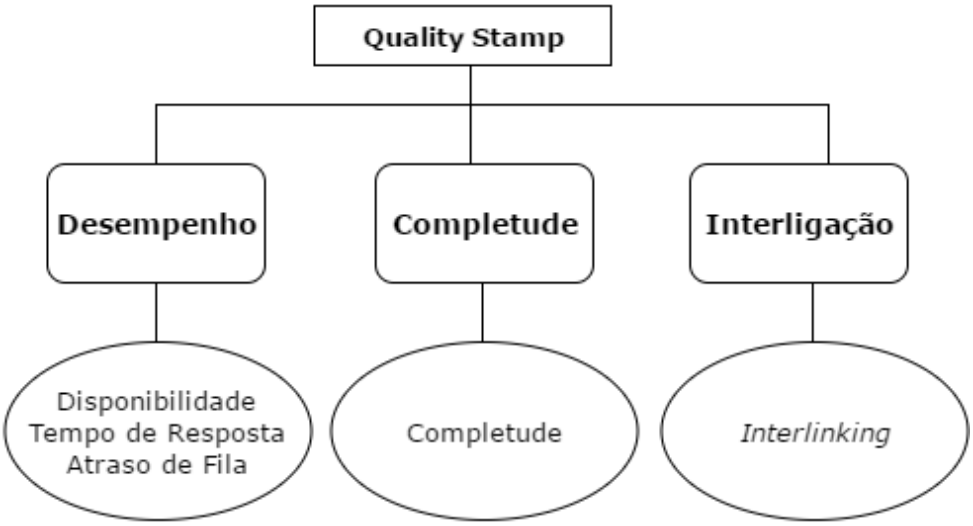
$$Rel(s_i) = \frac{(CE(s_i).CD(s_i)) + C(s_i)}{2} \tag{3.1}$$

3.1.4 Uma Abordagem para Avaliação de *Linked Datasets* para Aplicações de Domínio Específico

No trabalho proposto por Sarinho (2014), o enfoque principal é a avaliação de fontes de dados *Linked Data* para uma aplicação específica. Motivado pelo crescimento da *Web* de Dados, que por consequência tem possibilitado uma série de novas aplicações que utilizam fontes de dados *Linked Data*, o autor propõe a utilização de critérios de QI como maneira de identificar as melhores fontes de dados para uma aplicação de domínio específico.

A sua abordagem, denominada *Quality Stamp*, consiste em avaliar a qualidade das fontes de dados utilizando cinco critérios de QI, cujo objetivo é realizar a avaliação considerando três características das fontes de dados *Linked Data*, são elas: (i) desempenho, (ii) completude e (iii) interligação (Figura 3.2). Em seguida, é gerada uma medida única de qualidade para classificar as fontes de dados.

Figura 3.2: Características e Critérios de QI avaliados pela abordagem *Quality Stamp*



Fonte: Adaptado de Sarinho (2014)

A primeira característica avaliada tem como objetivo medir o desempenho de uma fonte de dados para responder solicitações das aplicações. Os critérios adotados foram **disponibili-**

dade, tempo de resposta e atraso de fila.

Para medir o critério de disponibilidade o autor implementou um *crawler* que de forma periódica realizava uma busca, na *Web* de Dados, por novas fontes de dados e as catalogava com o objetivo de monitorar a sua disponibilidade por meio de uma série de testes executados periodicamente. A métrica utilizada para disponibilidade consiste em uma requisição *ASK SPARQL* padrão que retorna *true* quando a fonte de dados está disponível e *false* caso contrário. O valor para o critério de disponibilidade consiste no percentual das requisições que responderam de forma positiva ao teste.

De semelhante modo, o tempo de resposta também foi monitorado. A métrica utilizada para esse critério consiste em uma requisição *ASK SPARQL* padrão, onde é capturado o tempo que foi gasto para responder tal requisição. O valor para o critério tempo de resposta consiste na somatória de todos os tempos obtidos por meio de uma série de testes executados periodicamente dividido pela quantidade total de requisições realizadas.

O critério atraso de fila é semelhante ao de tempo de resposta. O que diferencia um do outro é que esse critério tem como objetivo verificar o comportamento de uma fonte de dados quando recebe múltiplas requisições em um período de tempo curto. Para medir esse critério é realizada uma série de testes. Cada teste corresponde a uma quantidade x de requisições (também em forma de *ASK SPARQL* padrão[1]) que são feitas simultaneamente.

O valor para esse critério será uma porcentagem de variação de quanto a média dos tempos de respostas de cada teste difere do tempo de resposta de uma única requisição. Ao final, será efetuada a média ponderada entre os valores dos n testes.

A segunda característica avaliada foi a de Completude, cujo objetivo é identificar o quanto uma fonte de dados *Linked data* é completa em relação aos requisitos da aplicação. Nesse sentido, os requisitos da aplicação são descritos por um conjunto de consultas *SPARQL* $Q = \{q_1, ..., q_n\}$. O critério adotado para avaliar essa característica foi o de **Completude**. Por ser um critério abrangente, ele foi mensurado por meio de duas subclassificações: completude de Esquema e de dados.

A métrica utilizada para avaliar a completude do esquema consiste em verificar a existência dos padrões de triplas sem literais em cada fonte de dados *Linked Data* candidata à aplicação. Já a completude de dados é mensurada por meio de dois outros critérios propostos: a completude de literal e a completude de instância. A primeira verifica a existência de padrões de triplas com literal e a segunda verifica a existência dos recursos que representam instâncias de classes.

Sendo assim, o critério de completude (de forma geral) é mensurado por meio de uma média ponderada do critério de completude de dado e do esquema. É importante salientar que, os padrões de triplas são extraídos com base no conjunto de consultas Q que representam os requisitos da aplicação.

Por último, a terceira característica avaliada foi a de Interligação. O critério proposto para essa característica foi ***interlinking***, cujo objetivo é identificar o grau de interligação que

[1] Uma consulta *ASK SPARQL* pode ser definida da seguinte maneira: *ASK WHERE { ?s ?p ?o}*.

uma fonte de dados *Linked Data* possui com outras. Em outras palavras, quanto mais ligações uma fonte de dados possuir, mais ela pode fornecer informações relacionadas aos recursos que estão presentes nela.

A métrica utilizada para mensurar esse critério consiste na divisão entre o total de triplas com predicados de *interlinking* pelo total de triplas contidas na fonte de dados. O total de triplas pode ser encontrado utilizando uma consulta *SPARQL* por meio do operador *count*. Já para mensurar o total de triplas que possuem propriedades de *interlinking* podem ser utilizadas as seguintes consultas *SPARQL* (i) *SELECT distinct count(*) WHERE { ?s owl:equivalentClass ?o }*, (ii) *SELECT distinct count(*) WHERE {?s owl:equivalentProperty ?o}* e (iii) *SELECT distinct count(*) WHERE { ?s owl:sameAs ?o }*.

Uma vez que os cinco critérios foram mensurados, a abordagem gera uma medida única de qualidade. Essa medida é utilizada para classificar as fontes de dados candidatas à aplicação. Por utilizar critérios de QI de custos como tempo de Resposta e atraso de Fila é necessário realizar a normalização dos valores.

Com relação à normalização dos valores, o método utilizado foi o *SAW* (ZHU; BUCHMANN, 2002). Uma vez que os valores estão normalizados, a medida única de qualidade é calculada para cada fonte de dados por meio de uma média ponderada. Ao final da abordagem, é possível obter uma classificação de todas as fontes de dados que foram avaliadas, onde a que possuir o maior valor em sua medida única de qualidade representa aquela que melhor se adequa aos requisitos da aplicação.

3.1.5 Caracterizando e Selecionando Fontes de Dados *Fresh*

Analisar coletivamente múltiplas fontes de dados pode melhorar significantemente o valor da informação gerada. No entanto, o custo computacional para adquirir e integrar todos os dados disponíveis pode ser elevado. Nesse cenário, o trabalho de Rekatsinas, Dong e Srivastava (2014) propõe uma abordagem para selecionar um subconjunto ótimo, dentre um conjunto de fontes de dados, para realizar a integração.

O referido trabalho é uma extensão do que foi proposto em Dong, Saha e Srivastava (2012), porém, o seu diferencial é que o aspecto dinâmico das fontes é considerado. Nesse contexto, uma fonte de dados dinâmica é aquela cujo conteúdo muda ao longo do tempo.

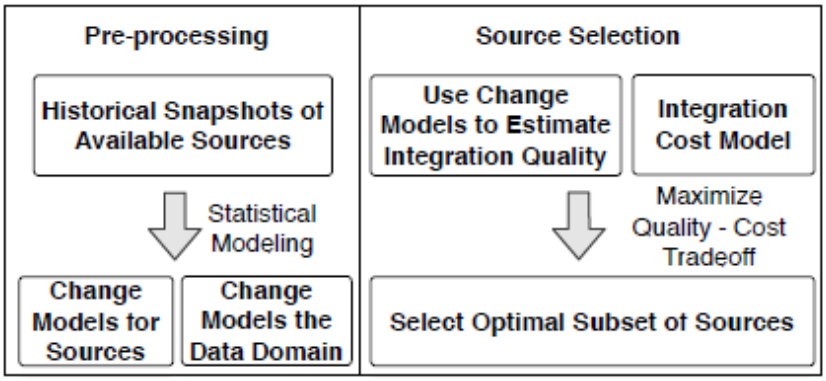
Para tratar a questão da dinamicidade das fontes no processo de seleção, os autores enfrentaram alguns desafios, dentre eles: (i) nem sempre as fontes que tem uma alta frequência de atualização são aquelas que oferecem os melhores dados; e (ii) a qualidade de uma fonte de dados pode mudar ao longo do tempo.

Para tratar esses desafios foi proposto um *framework*, na qual foi dividido em dois componentes (Figura 3.3). O primeiro componente, chamado de pré-processamento, avalia as fontes de dados de forma contínua utilizando alguns critérios de QI, como **cobertura e acurácia**, e utiliza modelos estatísticos para descrever os padrões de atualização dos dados e a frequência

de atualização da fonte.

O segundo componente, chamado de seleção de fontes, utiliza uma função de ganho (definida anteriormente) e os resultados que foram encontrados pelo componente de pré-processamento para estimar o comportamento da fonte de dados em um tempo futuro, e a partir disso selecionar o melhor conjunto de fontes a serem integradas em um instante de tempo t.

Figura 3.3: *Framework* para Seleção de Fontes de Dados



Fonte: Rekatsinas, Dong e Srivastava (2014)

Para validar a abordagem proposta foram utilizados dados reais, em dois cenários distintos, e dados sintéticos. Os experimentos mostraram que a proposta foi eficiente e conseguiu alcançar os objetivos que foram elencados pelos autores.

3.2 Geração e Utilização de Metadados das Fontes de Dados

Fornecer metadados é um aspecto importante quando se publica uma fonte de dados. Os metadados fornecem descrições de uma fonte de dados e podem ser utilizados tanto por consumidores de dados quanto por máquinas. Nesse contexto, selecionamos alguns trabalhos oferecidos na literatura cujo foco é gerar e/ou utilizar metadados para os mais diversos fins. A seguir, uma resumo sobre cada um deles é apresentado.

3.2.1 Usando Metadados de Qualidade para Seleção e *Ranking* de Fontes de Dados

O trabalho de Mihaila, Raschid e Vidal (2000) propõe o uso de metadados de qualidade para auxiliar na tarefa de seleção de fontes de dados. Localizar dados relevantes para um problema particular ainda é uma tarefa que demanda tempo, devido à vasta quantidade de fontes de dados disponíveis na *Web*. Dessa forma, os autores justificam que uma forma de minimizar esse custo é mantendo metadados que ajudem a descrever as fontes de dados bem como a sua qualidade.

Sendo assim, o enfoque principal do trabalho em questão consiste em definir um modelo no qual os metadados possam ser representados. De uma maneira geral, o modelo é composto por um conjunto de informações que descrevem o conteúdo da fonte de dados, tais como seus atributos e relacionamentos e por informações de qualidade. Nesse trabalho as informações de qualidade são denominadas de parâmetros. Sendo assim, os parâmetros incorporados ao modelo foram: **completude, atualidade, frequência de atualização e granularidade**.

Em nenhum momento durante o trabalho os autores destacaram como os parâmetros vão ser medidos. Isso se deve ao fato de que é assumido que esses valores são fornecidos pelos proprietários das fontes de dados.

Como parte da contribuição do trabalho, além da definição do modelo, uma linguagem para consultar as informações contidas nos metadados foi proposta. A Figura 3.4 mostra um exemplo dos metadados de qualidade gerados, baseado no modelo proposto do referido trabalho, para uma fonte de dados. A Figura 3.5 mostra um exemplo de como os metadados são consultados. No exemplo, a consulta busca selecionar as duas melhores fontes de dados que forneçam informações sobre a temperatura em Toronto para o ano corrente. Os critérios de qualidade ou parâmetros para a seleção foram *completude e granularidade*. As fontes selecionadas devem ter pelo menos 60% dos dados completos e a granularidade da informação ser perto de 1 hora.

Figura 3.4: Metadados de Qualidade

```
<?xml version="1.0"?>
<ws>
 <source>
  ... (source connection information)
  <metadata>
      <scqd type = "Air">
         <domain attr="location"
                 domtype="enumeration"
                 values="Halifax Montreal Toronto"/>
         <domain attr="time" domtype="range"
                 minvalue="Jan 1, 1990"
                 maxvalue="Dec 31, 2000"/>
         <qod>
           <subdomain attr="location"
                 domtype="enumeration"
                 values="Toronto"/>
           <subdomain attr="time" domtype="range"
                 minvalue="Jan 1, 2000"
                 maxvalue="Dec 31, 2000"/>
           <dimension name="completeness"
                      value="1.0" />
           <dimension name="last_updated"
                      value="Mar 1, 2000" />
           <dimension name="granularity"
                     attr="time" value="1 hour" />
           <measure attrs="temperature pressure"/>
         </qod>
      </scqd>
      ...
  </metadata>
 </source>
</ws>
```

Fonte: Mihaila, Raschid e Vidal (2000)

Figura 3.5: Exemplo - Consulta aos Metadados de Qualidade

```
select  best 2 s
from    Source s,
        Scqd c in s.scqds,
        Qod q in c.qods
where   c.type = "Air"
        and q.soma ≥ {"temperature"}
        and q.lcd ≥ [(city,{Toronto}),(time,CurrentYear)]
        and q.completeness > 0.6
        and q.granularity close to "1 hour";
```

Fonte: Mihaila, Raschid e Vidal (2000)

3.2.2 Um Perfil de Dados Urbanos

Motivados pelo grande volume de dados urbanos que estão sendo disponibilizados em diversos portais de dados abertos, o trabalho de Ribeiro et al. (2015) propõe uma ferramenta chamada UrbanProfiler, cujo objetivo é extrair automaticamente, por meio de tarefas de *data profiling* metadados da fonte de dados.

Encontrar fontes de dados relevantes para um determinado uso pode ser um problema a ser enfrentado, devido ao grande número de fontes de dados disponíveis e também ao fato de que os portais de dados abertos, que disponibilizam tais fontes, limitam a busca pelas fontes por palavras-chaves, por exemplo, considerando apenas uma descrição textual.

Nesse contexto, o *UrbanProfile* é responsável por analisar as fontes de dados e extrair metadados que descrevam o seu conteúdo, tais como: descrição de tipos de atributos, distribuição de valores, informações geográficas, entre outros. O objetivo é utilizar esses metadados para permitir que consultas possam ser enriquecidas com outros elementos e ajudar os usuários a identificar as fontes de dados que necessitam para um determinado uso.

O *UrbanProfile* possui dois componentes principais: (i) Extrator de Metadados e (ii) Análise da Fonte de Dados. O módulo extrator é responsável por acessar os portais de dados e realizar o *download* dos dados da fonte. Uma vez que o *download* é realizado, a fonte de dados passa pelo processo de verificação do seu conteúdo para que os metadados possam ser gerados.

Os metadados são gerados considerando duas perspectivas: a fonte de dados e o seu conteúdo (Figura 3.6). Considerando a fonte de dados, são extraídas informações que vão desde aspectos mais gerais como: número de linhas e colunas, até informações que detalham o volume de dados e data da última atualização. Com relação ao conteúdo, informações de distribuições de valores são extraídas para cada coluna da fonte de dados como: número total de valores únicos e faltantes, valores mais frequentes, entre outros.

Figura 3.6: Lista de metadados extraídos - *UrbanProfile*

Target	Group	Metadata
Dataset	General	Number of rows, columns, values and missing values
Dataset	Provided	Name, description, author, owner, category, update frequency and time of creation and last update
Dataset	Process	Input file size, total memory used, time (begin & end) and processing time
Column	General (all types)	Number of total, unique and missing values, and most frequent value
Column	Temporal	Minimum (min) and Maximum (max)
Column	Numeric	Min, max, mean and std of values
Column	Spatial	All unique values and Bounding Box
Column	Textual	Min, max, mean and std of text length

Fonte: Ribeiro et al. (2015)

O componente de análise auxilia a visualização dos metadados que foram gerados, que podem ser apresentados em três formas: mapas (para informações geolocalizadas), gráficos e tabelas. Esse componente também permite que diversas fontes possam ser comparadas por meio dos seus respectivos metadados.

Como estudo de caso, os autores utilizaram fontes de dados do portal de Dados NYC *Open Data*[2] e por meio dos metadados que foram gerados fizeram diversas análises mostrando a viabilidade da ferramenta.

3.2.3 *Linked Data Profiling*

No contexto de *Linked Data*, o trabalho de Abele (2016) apresenta uma abordagem para identificar o domínio de um conjunto de dados *Linked Data* por meio de seu conteúdo e metadados. A iniciativa *Linked Open Data* (LOD) ganhou uma grande visibilidade desde a sua criação. O crescimento rápido do número de conjunto de dados disponíveis revela o interesse dos produtores de dados em publicar seus dados nesse padrão.

Apesar do quarto princípio do padrão *Linked Data* ser a inclusão de *links* para outros conjuntos de dados que possuem recursos semelhantes, com o objetivo de descobrir mais conjuntos de dados que possam fornecer informações associadas, a grande maioria dos conjuntos de dados não é ligada.

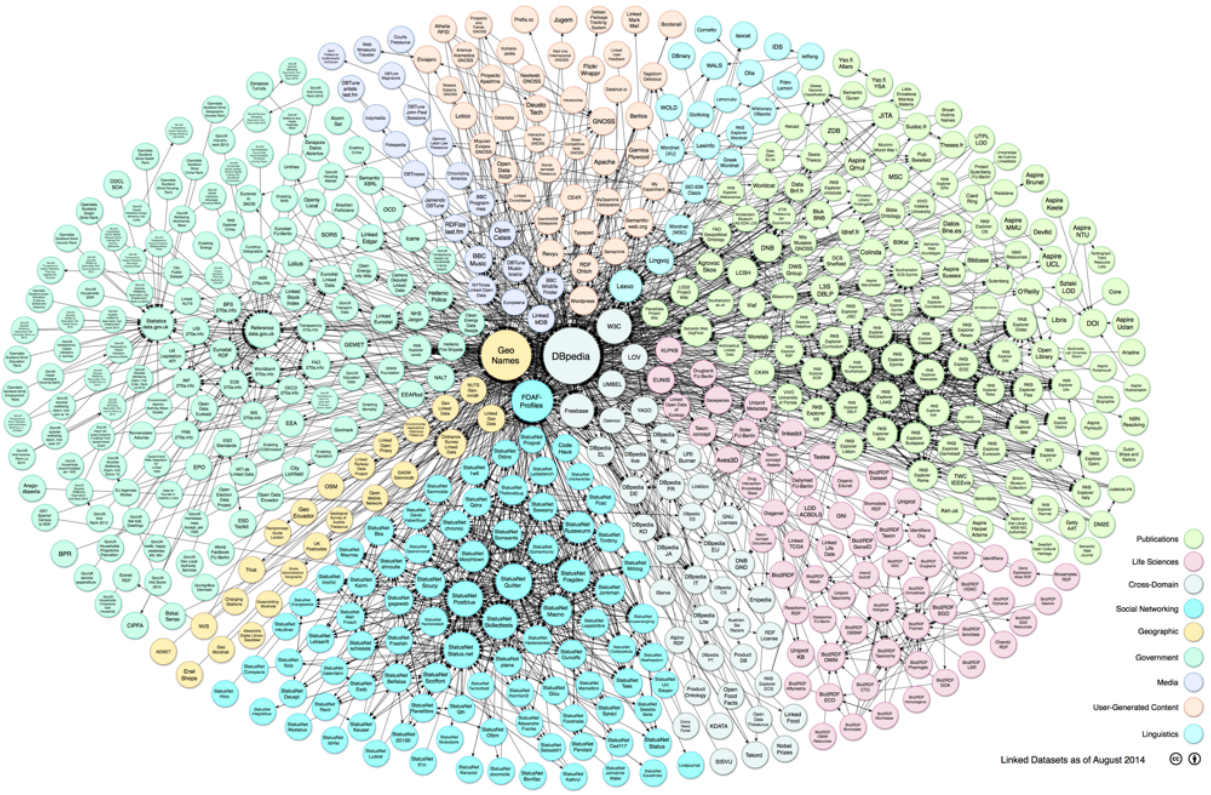
O autor observou que dos 10.632 conjuntos de dados disponíveis no DataHub[3] apenas 1.207[4] possuiam ligações com outros conjuntos de dados. O autor acredita que isso se deve ao fato de que os novos produtores de dados não possuem informações suficientes dos conjuntos de dados já existentes, fazendo com que esse problema se prolonge.

[2]https://nycopendata.socrata.com/
[3]https://datahub.io
[4]Consulta realizada em 22/11/15

Como forma de resolver esse problema, o referido trabalho propõe duas abordagens. A primeira abordagem realiza uma análise e fornece informações detalhadas dos recursos contidos no conjunto de dados, fornecendo um conjunto de metadados, tais como: estatísticas sobre frequências e distribuições de sujeitos, predicados e objetos, uma lista dos diferentes tipos de dados usados para literais, e uma lista de vocabulários utilizados. A segunda abordagem, por meio dos metadados gerados, sugere quais conjuntos de dados existentes podem ser ligados ao conjunto de dados que um produtor de dados quer publicar.

O trabalho foi validado por meio de um *baseline* criado de acordo com o diagrama do LOD *cloud*[5] (Figura 3.7) publicado em 2014. O diagrama contém 342 conjuntos de dados ativos em 9 domínio de dados diferentes. Os experimentos realizados comprovaram que aplicando as duas abordagens propostas em tais conjuntos de dados foi possível obter uma precisão considerável acerca da classificação dos conjuntos de dados utilizando os metadados gerados, alcançando assim o objetivo esperado.

Figura 3.7: Diagrama do LOD *cloud*



Fonte: http://lod-cloud.net/

3.3 Discussões

Com exceção de Batista (2008), os demais trabalhos apresentados na Seção 3.1 abordam o problema de seleção de fontes de dados. De uma maneira geral, esses trabalhos consideram o

[5]http://lod-cloud.net/

uso de critérios de QI para avaliar as fontes de dados e propõem uma abordagem para seleção baseada na qualidade das fontes de dados.

Em Batista (2008), apesar do foco ser avaliação do esquema integrado, o que foge um pouco do escopo principal desta dissertação, é um trabalho que merece ser destacado, uma vez que é bem fundamentado no que se refere a conceitos relacionados a QI. O trabalho não se limita apenas na avaliação da qualidade do esquema integrado, mas, expande sua pesquisa a outros elementos de um sistema de integração, como as fontes de dados propriamente ditas. Especificamente, essa contribuição foi bastante útil para esta pesquisa.

A abordagem proposta por Xian et al. (2009) apesar de ter um processo de avaliação bem definido, possui algumas características que podem ser cruciais quando avaliamos qualidade. Os autores propõem avaliar os critérios de Completude e Corretude dos Dados considerando apenas uma amostra dos dados. No contexto do referido trabalho, é considerado que as fontes de dados geram um grande volume de dados, por isso o processo de extrair e analisar todos os dados das fontes pode ser custoso. Esse tipo de abordagem pode se mostrar ineficiente, pois os valores atribuídos a esses critérios podem não condizer com a realidade, uma vez que os dados coletados para definir a amostra podem levar a valores muito além ou aquém da realidade.

As abordagens propostas por Loscio et al. (2012) e Sarinho (2014) são bastante semelhantes e também têm seus processos de avaliação bem definidos. No entanto, para avaliar a qualidade das fontes de dados são considerados os requisitos da aplicação, o que torna a avaliação da qualidade muito específica. O problema de uma avaliação específica é que os resultados não podem ser reutilizados em outros cenários.

O trabalho de Rekatsinas, Dong e Srivastava (2014) é o único, dentre os que citamos acima, que considera o aspecto dinâmico das fontes. É uma abordagem com um processo de avaliação bem definido que a partir de uma avaliação contínua da qualidade das fontes de dados consegue, por meio de análises estatísticas, estimar o comportamento das fontes de dados em tempo futuro e descobrir um melhor conjunto de fontes a serem integradas ao longo do tempo.

Com relação aos trabalhos apresentados na Seção 3.2, todos, com exceção do de Abele (2016), abordam o problema de seleção de fontes de dados propondo como solução o uso de metadados das fontes de dados.

Em Mihaila, Raschid e Vidal (2000) os autores assumem que as fontes de dados oferecem metadados que descrevem o seu conteúdo e informações acerca da sua qualidade para propor um modelo de dados para representar tais metadados e uma linguagem para consultar esses metadados. O objetivo é que esses metadados possam enriquecer uma consulta e auxiliar no processo de selecionar melhores fontes de dados para uma necessidade específica.

Os trabalhos de Ribeiro et al. (2015) e Abele (2016) utilizam técnicas de *Data Profiling* para extrair informações detalhadas das fontes de dados e gerar metadados que podem ser associados às fontes. O primeiro, trata o problema da seleção de fontes de dados e propõe o uso dos metadados gerados para o enriquecimento de consultas e para auxiliar os usuários a encontrar fontes de dados mais relevantes para sua necessidade. O segundo, considera o contexto

de *Linked Data* e define uma abordagem que utiliza os metadados gerados para recomendações de fontes de dados.

O Quadro 3.1 apresenta uma sumarização das principais características identificadas nos trabalhos relacionados

3.4 Considerações

Neste capítulo foram analisados os trabalhos relacionados ao tema desta dissertação. Ao final, foi realizada uma breve discussão entre os trabalhos, agrupando as principais características de cada abordagem em uma tabela comparativa. Foram levantados os pontos mais relevantes de cada um com o objetivo de identificar o que já existe com relação ao tema central desta pesquisa. Foi possível observar que poucos trabalhos consideram o aspecto dinâmico das fontes no processo de avaliação da qualidade e nenhum desses trabalhos gera metadados de qualidade para fontes de dados. O próximo capítulo apresenta a proposta central desta dissertação.

Quadro 3.1: Sumarização - Trabalhos Relacionados

Trabalho	Estratégia de Avaliação	Contexto da Avaliação	Geração de Metadados de Qualidade
Batista (2008)	Pontual	Específico	Não
Xian et al. (2009)	Pontual	Geral	Não
Loscio et al. (2012)	Pontual	Específico	Não
Sarinho (2014)	Mista*	Específico	Não
Rekatsinas, Dong e Srivastava (2014)	Contínua	Geral	Não
Mihaila, Raschid e Vidal (2000)	Não relatado**	Não relatado**	Não
Ribeiro et al. (2015)	Pontual	Geral	Parcial***
Abele (2016)	Pontual	Geral	Não

Fonte: O Autor (2016)

Legendas:

* O trabalho avalia apenas os critérios disponibilidade, tempo de resposta e atraso de fila de maneira contínua.

** O trabalho não detalha como os valores para os critérios de qualidade foram atribuídos e nem como as fontes de dados foram avaliadas.

*** Dentre os metadados que são gerados por esse trabalho, a informação de quantidade de valores faltantes é uma delas e pode ser considerado como um elemento de qualidade.

4

Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas

Neste capítulo iremos especificar o processo para geração de um Perfil de Qualidade para fontes de dados dinâmicas. Inicialmente, abordaremos o conceito do Perfil de Qualidade, em seguida algumas definições preliminares e a especificação dos critérios de qualidade selecionados para compor o Perfil de Qualidade serão introduzidas para auxiliar no entendimento do processo que será descrito posteriormente.

Os critérios de QI selecionados para compor o Perfil de Qualidade serão descritos na Seção 4.2 e a especificação do processo para geração do Perfil de Qualidade será detalhada na Seção 4.3. Ao fim deste capítulo, na Seção 4.4, comparamos este trabalho com os trabalhos relacionados, apresentados no capítulo anterior.

4.1 Perfil de Qualidade

O Perfil de Qualidade consiste de um conjunto de metadados que descrevem a qualidade de uma fonte de dados em termos de um conjunto de critérios de QI.

Inspirado no conceito de *Data Profiling* (NAUMANN, 2014), que é o processo de examinar os dados disponíveis em uma fonte de dados a fim de coletar estatísticas e informações sobre o mesmo, o Perfil de Qualidade busca expandir esse conceito enriquecendo a descrição de uma fonte de dados associando informações acerca de sua qualidade. Para isso, levamos em consideração dois aspectos: (i) qual a importância da informação de qualidade de uma fonte de dados para seus utilizadores? e (ii) como avaliar a qualidade de uma fonte de dados?

No Capítulo 3 analisamos alguns trabalhos que avaliam a qualidade das fontes de dados e utilizam essas informações para diversos fins, como, determinar a adequação de uma fonte de dados para um determinado uso e até mesmo no processo de selecionar dentre um conjunto de fontes de dados quais aquelas que seriam mais relevantes para uma determinada aplicação. Trabalhos como os que apresentamos justificam a necessidade de se avaliar a qualidade das fontes de dados e manter associados a elas metadados de qualidade.

O processo de gerar metadados de qualidade está intimamente ligado à avaliação da qualidade de uma fonte de dados. Especificamente, neste trabalho estamos abordando o problema de avaliação da qualidade de fontes de dados dinâmicas.

Por estarmos considerando o aspecto dinâmico das fontes, adotamos uma estratégia de avaliação contínua. Isso implica dizer que o Perfil de Qualidade poderá ter suas informações atualizadas de forma periódica, segundo a frequência de atualização da fonte de dados, para que possa entregar aos usuários uma informação mais precisa da qualidade da fonte.

O Perfil de Qualidade se apresenta como um facilitador para os consumidores e produtores de dados, uma vez que o seu uso pode ser aproveitado para diversos fins. Destacamos para os consumidores de dados: (i) avaliar se a fonte de dados se adequa para seu uso; e (ii) auxiliar no processo de seleção de fontes para uma atividade específica. Para os produtores, será possível acompanhar a evolução da qualidade da fonte por meio de uma análise do perfil.

As informações oferecidas pelo Perfil de Qualidade também podem ser utilizadas para atividades mais complexas, como é o caso da adaptação de uma consulta sensível à qualidade. Nesse contexto, uma consulta sensível à qualidade é aquela que inclui as preferências do usuário para a qualidade dos dados nas respostas da consultas a partir de múltiplas fontes de dados (YEGANEH et al., 2009).

Antes de apresentarmos a formalização do problema que estamos tratando, alguns conceitos serão definidos.

4.1.1 Definições Preliminares

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

Definição 4.1. Conjunto de Critérios de Qualidade - $Q = \{Q_1, ..., Q_n\}$, denota um conjunto de critérios de qualidade, tal que cada $Q_i \in Q$ representa um critério utilizado para avaliar e medir um aspecto específico da qualidade de uma fonte de dados f.

Definição 4.2. Conjunto de Valores dos Critérios de Qualidade - O conjunto CQ consiste de um conjunto de pares, $\{(Q_1, v_1), ..., (Q_n, v_n)\}$, 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 4.3. 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 4.4. 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 4.5. 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.

4.1.2 Definição do Problema

Seja f uma fonte de dados dinâmica que sofre modificações em uma frequência λ e $Q = \{Q_1, ..., Q_n\}$ um conjunto de critérios de QI. Como avaliar a qualidade de f com relação a Q e gerar o seu respectivo $PQ(f)$ considerando o seu aspecto dinâmico?

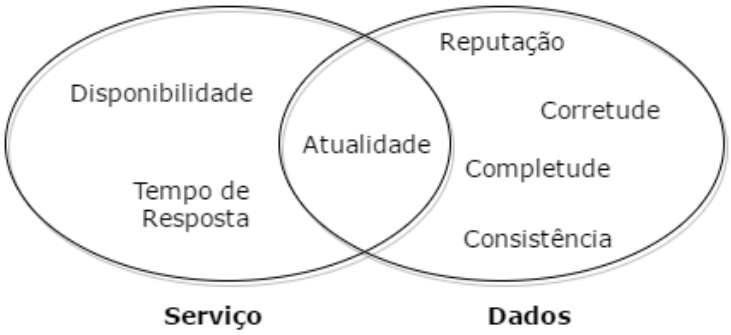
4.2 Critérios de Qualidade e Métricas de Avaliação

Uma avaliação da qualidade é concebida por meio de valores agregados de múltiplos critérios de qualidade que são relevantes aos consumidores da informação (ZAVERI et al., 2012). Logo, para cada critério de qualidade será associada uma ou mais métricas de avaliação.

Com relação aos critérios de qualidade, a literatura oferece uma vasta quantidade de trabalhos que propõem e descrevem cada um deles (WANG; STRONG, 1996; NAUMANN; ROLKER, 2000; ZHU; BUCHMANN, 2002).

Com base no levantamento da literatura, apresentado na Seção 2.1.1, identificamos critérios de QI que podem ser relevantes para avaliar a qualidade de uma fonte de dados. Nesse sentido, propomos uma classificação para tais critérios, apresentados na Figura 4.1.

Figura 4.1: Classificação dos Critérios de QI Relevantes para Avaliação de uma Fonte de Dados



Fonte: O Autor (2016)

Os critérios foram divididos em duas (2) categorias: (i) serviço; e (ii) dados. Os critérios relacionados à primeira categoria dizem respeito a critérios que avaliam a fonte de dados considerando a qualidade do serviço que é provido por ela. Nessa categoria a avaliação contempla pontos específicos ao comportamento da fonte como um serviço, tais como: disponibilidade e tempo de resposta. Os critérios relacionados à segunda categoria dizem respeito a critérios que avaliam especificamente o conteúdo da fonte de dados, ou seja, os dados propriamente dito. É

importante salientar que, as definições para tais critérios e suas respectivas métricas já foram apresentados na Seção 2.1.2.

Especificamente, neste trabalho, iremos ilustrar a estratégia de avaliação contínua da qualidade de uma fonte de dados e a geração do seu respectivo Perfil de Qualidade, considerando a categoria dados e apenas os critérios de qualidade: **completude e corretude**. Os critérios foram escolhidos por serem dois aspectos pertinentes, que podem ser avaliados independentemente 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 qualidade 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. A seguir, iremos descrever os critérios selecionados, especificando como eles serão mensurados. As métricas aplicadas são todas objetivas, o que deixa a avaliação da qualidade independente do usuário.

4.2.1 Completude

O critério completude indica o grau em que os dados de uma fonte são completos para uma determinada tarefa (WANG; STRONG, 1996). Em Pipino, Lee e Wang (2002), os autores propõem mensurar a completude considerando tanto o esquema quanto os dados da fonte. O esquema de uma fonte é dito completo se, ao compararmos o esquema da fonte com um esquema de referência, todos os conceitos do esquema de referência estiverem presentes no esquema da fonte.

Neste trabalho, estamos interessados apenas na completude dos dados, uma vez que não consideramos a existência de um esquema de referência. Com relação aos dados, o grau de Completude de uma fonte pode ser estimado com base nos valores nulos e faltantes, por meio das métricas de Densidade e Cobertura (NAUMANN; FREYTAG, 2000), respectivamente.

Definição 4.6. Densidade - A densidade 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, os quais são descritos pelo conjunto de atributos $A = \{a_1, ..., a_n\}$, onde cada atributo $a_i \in A$ possui 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 atributo $a_i \in A$, denotada por $d(a_i)$.

A densidade de um atributo a_i é dada pelo percentual de valores não nulos contidos em $V(a_i)$ (Equação 4.1).

$$d(a_i) = \frac{\alpha}{\beta} \tag{4.1}$$

Onde,

α, é a quantidade de valores não nulos contidos em $V(a_i)$;

β, é a quantidade total de valores contidos em $V(a_i)$.

Ao final, a densidade da fonte $D(f)$ é dada pelo somatório de todos os valores de $d(a_i) \in A$, dividido pelo número de atributos contidos em A, representado por $|A|$ (Equação 4.2).

$$D(f) = \frac{\sum_{i=1}^{|A|} d(a_i)}{|A|} \tag{4.2}$$

Onde,
$d(a_i)$, é o valor da densidade do atributo a_i;
$|A|$, é o total de atributos da fonte de dados.

Definição 4.7. Cobertura - A cobertura de uma fonte de dados f, denotada por $C(f)$, é dada pelo quantidade de instâncias contidas em f com relação à quantidade de instâncias que é esperada em f (Equação 4.3).

$$C(f) = \frac{|I|}{|W|} \tag{4.3}$$

Onde,
$|I|$, é o total de instâncias contidas na fonte de dados;
$|W|$, é o total de instâncias que é esperada na fonte de dados.

A métrica utilizada para avaliar a cobertura foi inspirada e adaptada do trabalho de Naumann e Freytag (2000). Porém, em alguns domínios nem sempre é possível determinar o total de instâncias esperadas em uma fonte de dados. Em situações que não seja possível saber o valor de $|W|$ para f, pode-se assumir a totalidade do valor para o critério em questão, ou seja, $C(f) = 1$ ou avaliar a completude com base apenas na densidade.

Definição 4.8. Completude - A completude de uma fonte de dados f, denotada por $Completude(f)$, é dada pela combinação dos valores de Densidade e Cobertura (Equação 4.4).

$$Completude(f) = D(f) * C(f) \tag{4.4}$$

Para contextualizar a avaliação do critério completude, considere que estamos avaliando uma fonte de dados hipotética, denominada F, que possui características semelhantes àquela utilizada no exemplo motivacional (Seção 1.2). Considere também que tal fonte oferece a seguinte amostra de dados, apresentada na Tabela 4.1.

Tabela 4.1: Amostra de Dados - Exemplo

DATA_MEDICAO	TEMP_MED	TEMP_MAX	TEMP_MIN
19/08/2014 03:00:00	27	28,5	27,5
19/08/2014 12:00:00	28,5	28,3	58
19/08/2014 13:00:00	29	29,5	29,1
19/08/2014 14:00:00	29	10	28,7
19/08/2014 16:00:00	null	null	null
19/08/2014 17:00:00	25,5	26,5	24,7
19/08/2014 18:00:00	23,7	25	22,8
19/08/2014 19:00:00	22,7	null	22,4
19/08/2014 20:00:00	22,6	22,9	22,3
19/08/2014 21:00:00	22,6	22,9	22,3
19/08/2014 22:00:00	22,6	22,9	22,3
19/08/2014 23:00:00	23,1	23,6	22,5

Fonte: O Autor (2016)

Primeiramente, será necessário calcular a densidade de cada atributo que descreve o conjunto de dados analisados, nesse caso, aplicando a Equação 4.1, temos que:

$$d(DATA\_MEDICAO) = (\frac{12}{12}) = 1$$

$$d(TEMP\_MED) = (\frac{11}{12}) = 0.92$$

$$d(TEMP\_MAX) = (\frac{10}{12}) = 0.83$$

$$d(TEMP\_MIN) = (\frac{11}{12}) = 0.92$$

Ao final, aplicando a Equação 4.2, a Densidade de F será calculada com base no somatório de todas as densidades de seus atributos, calculados anteriormente. Considerando isso temos que:

$$D(F) = (\frac{1+0.92+0.83+0.92}{4}) = 0.9175$$

Em outras palavras isso quer dizer que, 91,75% dos valores contidos na amostra de dados avaliadas não são nulos.

Em seguida, a cobertura de F é calculada. Considere que quem fornece dados para F é um sensor, e espera-se que este sensor forneça uma nova instância para F a cada hora. Imagine que a amostra de dados apresentada na Tabela 4.1 é referente ao período de tempo de 24 horas (1 dia). Logo, espera-se que em F estejam contidas 24 instâncias, ou seja, $|W| = 24$. Considerando a amostra de dados que estamos avaliando (Tabela 4.1), teríamos que $|I| = 12$. Sendo assim, aplicando a Equação 4.3, temos que a cobertura de F é:

$$C(F) = (\frac{12}{24}) = 0.5$$

Em outras palavras, apenas 50% das instâncias que eram esperadas estão presentes em F.

Dado que a densidade e cobertura foram calculadas, o valor atribuído ao critério de completude, aplicando a Equação 4.4, é 0.46.

$$Completude(F) = 0.9175 * 0.5 = 0.4587$$

4.2.2 Corretude

O critério corretude indica o grau em que os dados contidos em uma fonte de dados estão livres de erro (WANG; STRONG, 1996; LOSCIO et al., 2012). Outras nomenclaturas, como Verificabilidade, também podem ser usadas para denotar a corretude dos dados (WANG; STRONG, 1996). 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 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.

Nesta dissertação, a fim de avaliar o critério corretude, propomos o uso de regras de validação, as quais são definidas de acordo com o domínio dos dados e que podem ser usadas para uma avaliação automática da corretude. Nesse sentido, assumimos que dado um domínio, possuímos uma base de regras que foram previamente definidas por um usuário 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 a qual é pertencente.

Definição 4.9. 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 a fonte de dados apresentada anteriormente (utilizada para ilustrar o critério completude), e a seguinte restrição do domínio de dados meteorológicos (domínio a qual pertence a fonte de dados especificada): *a temperatura mínima deve ser menor que a temperatura média, que por sua vez deve ser menor ou igual a temperatura máxima.*

Nesse caso, poderíamos aplicar a seguinte regra como forma de validação da restrição descrita acima:

$R_1 = \{temp\_min < temp\_med\ \&\&\ temp\_med \leq temp\_max\}$;

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.

Sendo assim, o valor para o critério de corretude de uma fonte de dados f pode ser obtido pela Equação 4.5. Primeiro, a partir da verificação das regras de validação, calcula-se o número de instâncias incorretas da fonte f. Em seguida, esse valor é divido pelo total de instâncias avaliadas em f, e o resultado é subtraído de um (1) para que seja encontrado a porcentagem de instâncias corretas da fonte f.

$$Corretude(f) = 1 - (\frac{\delta}{\beta}) \tag{4.5}$$

Onde,
δ, é o total de instâncias incorretas;
β, é o total de instâncias avaliadas.

Para exemplificar a avaliação do critério corretude, considere que estamos avaliando a amostra de dados apresentada na Tabela 4.1, utilizando o conjunto de regras $R = \{R_1\}$ definido anteriormente. Supondo que as instâncias contidas em f passaram pelo processo de validação, foram encontradas as seguintes instâncias com erros (Tabela 4.2)

Tabela 4.2: Instâncias Incorretas - Exemplo

DATA_MEDICAO	TEMP_MED	TEMP_MAX	TEMP_MIN
19/08/2014 12:00:00	28,5	28,3	**58**
19/08/2014 14:00:00	29	**10**	28,7

Fonte: O Autor (2016)

Sendo assim, o valor atributído ao critério de Corretude, aplicando a Equação 4.5, é $Corretude(F) = 0.83$. Em outras palavras, 83% das instâncias disponíveis em F estão corretas.

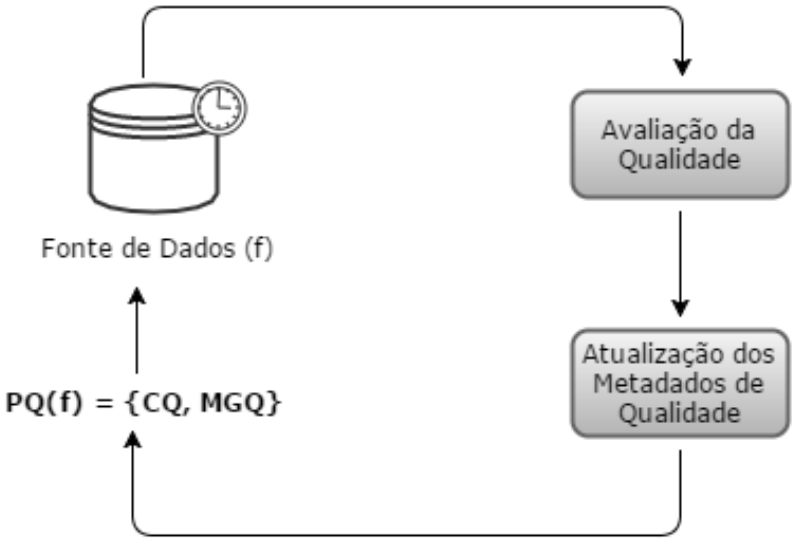
$$Corretude(F) = 1 - (\frac{2}{12}) = 0.83$$

4.3 Processo de 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 4.2 e descrito a seguir. A geração do Perfil de Qualidade segue uma estratégia de avaliação contínua da qualidade. Isso quer dizer que, uma vez que foi gerado, o Perfil de Qualidade poderá ter suas informações atualizadas de acordo com uma frequência de atualização previamente definida (λ).

O processo é divido em duas etapas que serão sintetizadas abaixo e detalhadas nas seções seguintes:

Figura 4.2: Visão Geral do Processo de Geração do Perfil de Qualidade



Fonte: O Autor (2016)

- **Etapa 1 - Avaliação da Qualidade** - Nessa etapa do processo, a fonte de dados é avaliada com base nos critérios de qualidade que foram definidos. Por estarmos considerando que as fontes de dados são dinâmicas, adotamos uma estratégia de avaliação contínua, ou seja, quando a fonte de dados é modificada uma nova avaliação deve ser realizada.

  Como vimos, quando discutimos o problema da avaliação em fonte de dados dinâmicas, o custo para reavaliar a fonte de dados por completo todas as vezes que a mesma sofre modificações pode ser bastante alto. Principalmente, quando um grande volume de dados é gerado ao longo do tempo, tornando assim esta solução inviável. Como contribuição para essa situação, a nossa estratégia de avaliação contínua segue um viés incremental. Detalhes serão apresentados na Seção 4.3.1.

- **Etapa 2 - Atualização dos Metadados** - Nessa etapa do processo, os valores dos critérios de qualidade são calculados e anotados no Perfil de Qualidade que será disponibilizado junto com a fonte de dados.

  Como contribuição nessa etapa, propomos uma maneira de calcular os valores dos critérios considerando o resultado da avaliação mais recente com resultado de avaliações passadas. O objetivo é fazer com que os valores atribuídos aos critérios de qualidade possam refletir com precisão o estado atual da fonte de dados sem ter que reavaliá-la por completo todas as vezes que a mesma sofrer modificações. Detalhes serão apresentados na Seção 4.3.2.

4.3.1 Etapa 1 - Avaliação da Qualidade

A primeira etapa do processo de geração do Perfil de Qualidade se concentra em avaliar a qualidade da fonte de dados por meio dos critérios de qualidade selecionados. Nessa etapa cada critério de qualidade $Q_i \in Q$ será mensurado conforme descrito na Seção 4.2.

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 de dados a serem avaliados pode ser muito grande. Para ilustrar essa situação considere a figura abaixo.

Figura 4.3: Exemplo - Instâncias Disponíveis em uma Fonte de Dados

Fonte: O Autor (2016)

A Figura 4.3 mostra uma ilustração de um determinado volume de dados disponíveis em uma fonte de dados. Nesse sentido, consideramos cada ponto presente no retângulo como uma instância da fonte de dados. A área total do retângulo, demarcada por A, representa o total de instâncias disponíveis na fonte em um dado momento.

Considerando que essa fonte de dados é dinâmica e que já foi avaliada outras vezes, a área do retângulo menor, demarcado por B, representa o total de novas instâncias que foram adicionadas na fonte de dados em um período que compreende a última avaliação da qualidade da fonte até o momento atual.

Especificamente, ao invés de avaliar as instâncias contidas na área demarcada por A, iremos considerar apenas as novas instâncias, ou seja, as instâncias contidas na área demarcada por B. Dessa forma, a amostra de dados é avaliada com base no conjunto Q de critérios de qualidade (Seção 4.2). Para cada $Q_i \in Q$ será atribuído um valor $\Delta(Q_i)$, esse valor representa o valor atribuído ao critério Q_i com base na amostra de dados utilizada na avaliação.

Entretanto, como foi discutido anteriormente, considerar apenas uma amostra isolada dos dados pode não ser suficiente para avaliar de maneira precisa a qualidade de uma fonte. Sendo assim, 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 próxima etapa do processo é responsável por realizar o cálculo dos valores dos critérios considerando os aspectos levantados acima.

4.3.2 Etapa 2 - Atualização dos Metadados de Qualidade

A segunda etapa do processo de Geração do Perfil de Qualidade para uma fonte de dados f consiste em atualizar os metadados de qualidade que fazem parte do $PQ(f)$. O primeiro passo dessa etapa é calcular os valores de v_i para cada um dos elementos do conjunto CQ e logo em seguida calcular a MGQ do $PQ(f)$.

Como discutimos anteriormente, a avaliação da qualidade realizada na etapa anterior só considera uma amostra dos dados. Para que os valores atribuídos aos critérios não sejam distantes do esperado, propomos um cálculo que leva em consideração o resultado da avaliação atual com resultados de avaliações realizadas em momentos anteriores.

Dessa forma, para cada $Q_i \in Q$ consideramos o valor de $\Delta(Q_i)$, calculado na etapa anterior, e o valor corrente do critério $Q_i.v_i$, contido no $PQ(f)$ para calcular seu novo valor v_i, denotado por $Q_i.v_{i(new)}$ (Equação 4.6). Os dois valores são combinados por w, que representa o peso que $\Delta(Q_i)$ possui na referida equação.

Uma observação a ser feita é que o valor de w equivale à proporção de instâncias que foram avaliadas e pode ser obtido automaticamente, dividindo o total de instâncias que foram avaliadas pelo total de instâncias contidas na fonte.

$$Q_i.v_{i_{(new)}} = w * \Delta(Q_i) + (1-w) * (Q_i.v_i) \tag{4.6}$$

Onde,
$Q_i.v_{i_{(new)}}$, é o novo valor do critério Q_i que será atualizado em $f.CQ \in PQ$;
$\Delta(Q_i)$, é o valor para o critério q_i obtido na avaliação realizada na etapa anterior;
w, corresponde ao peso que $\Delta(Q_i)$ assumirá no cálculo;
$Q_i.v_i$, é o valor corrente do critério q_i, armazenado no $PQ(f)$.

Após atualização dos valores dos critérios de qualidade da fonte de dados, o novo valor de $f.MGQ$ deverá ser calculado (Equação 4.7).

$$MGQ_{(new)} = \sum_{i=1}^{|Q|} w_i * Q_i.v_i \tag{4.7}$$

Onde,
$|Q|$, é o total de critérios de qualidade selecionados para ser utilizado na avaliação;
w_i, corresponde ao peso atribuído a um critério q_i;
$Q_i.v_i$, é o valor atual do critério q_i.

É importante salientar que a soma dos pesos de todos os critérios não pode ser maior que 1. Logo, o peso atribuído a um critério deve ser proporcional à 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.

O Quadro 4.1 apresenta uma sumarização das informações disponibilizadas no Perfil de Qualidade.

Quadro 4.1: Estrutura do Perfil de Qualidade

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

Fonte: O Autor (2016)

Destacamos que esta dissertação não contempla como os metadados serão representados e entregues aos usuários. No documento de Melhores Práticas para Publicação e Consumo de Dados na Web, disponibilizado pela W3C (LOSCIO; BURLE; CALEGARI, 2016), é sugerido que metadados de qualidade sejam representados em formato de metadados *machine readable*. Trabalhos que tratam essa questão já têm sido propostos (DEBATTISTA; LANGE; AUER, 2014).

Na Figura 4.4 apresentamos um exemplo do Perfil de Qualidade, com informações de completude, corretude e medida global, sendo representado no formato JSON[1].

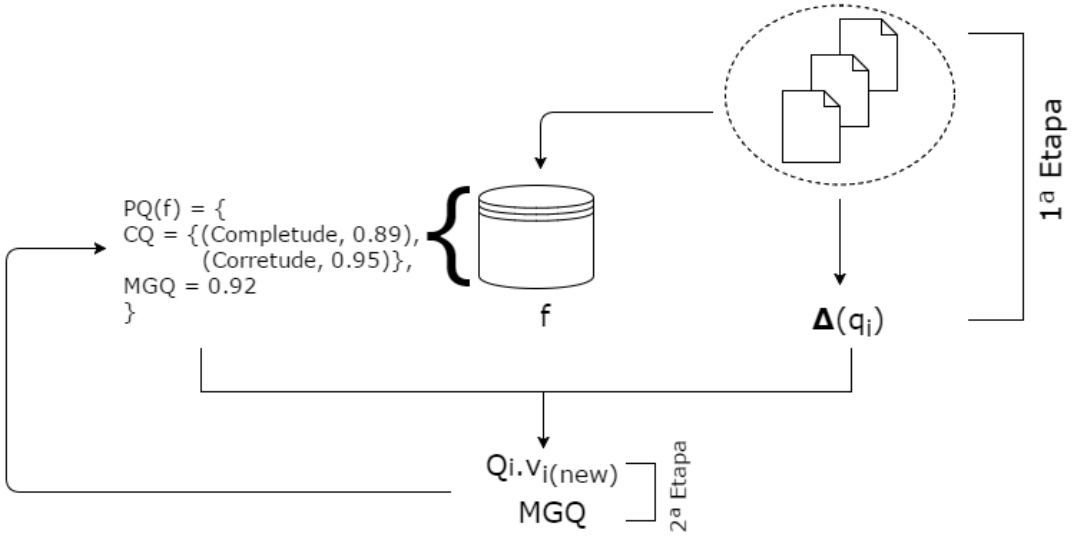
Figura 4.4: Exemplo do Perfil de Qualidade em JSON

```
{
   "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
}
```

Fonte: O Autor (2016)

[1]É um formato para intercâmbio de dados computacionais, formado por um subconjunto de notações de objeto de JavaScript

Figura 4.5: Exemplo - Processo de Geração do Perfil de Qualidade



Fonte: O Autor (2016)

4.3.3 Exemplo

Para ilustrar o processo de geração do Perfil de Qualidade, iremos apresentar um exemplo detalhando cada etapa descrita nas seções anteriores.

A Figura 4.5 mostra um exemplo de uma fonte de dados f que possui um comportamento dinâmico. Considere que a frequência de atualização do Perfil de Qualidade de f é diária e que f já passou pelo processo outras vezes, portanto já possui um Perfil de Qualidade gerado. Dessa forma, iremos demonstrar como o seu Perfil de Qualidade será modificado em função da evolução que f sofre ao longo do tempo.

Do lado esquerdo da figura é possível observar o Perfil de Qualidade atual de f, representado por $PQ(f)$. É possível observar também do lado direito, na parte superior, novas instâncias que foram adicionadas em f no intervalo de tempo que compreende a última avaliação até o instante de tempo atual. Considerando que tais modificações ocorreram em f, é necessário realizar uma nova avaliação da qualidade e atualizar o seu Perfil de Qualidade para que as informações contidas em seu Perfil possam refletir a modificação que foi realizada.

Utilizando a estratégia de avaliação proposta, apenas as novas instâncias serão avaliadas. Imagine que f possui 10.000 instâncias, sendo 1.000 delas as novas instâncias que iremos avaliar.

Supondo que foi realizada a avaliação da qualidade nas novas instâncias, utilizando os dois critérios apresentados na Seção 4.2 (1ª etapa do processo), hipoteticamente temos que: $\Delta(Q_1) = 0.80$ e $\Delta(Q_2) = 0.75$, onde Q_1 equivale ao critério completude e Q_2 ao critério corretude.

Uma vez que a etapa da avaliação foi realizada, é necessário calcular os novos valores dos critérios que serão atualizados no Perfil de Qualidade (2ª etapa do processo). Nesse caso, temos que recuperar os valores dos critérios armazenados em CQ. Sendo assim, temos que:

$CQ = \{(Q_1, 0.89), (Q_2, 0.95)\}$.

Para que os novos valores dos critérios sejam calculados e atualizados em CQ podemos utilizar a Equação 4.6. O cálculo será exemplificado abaixo.

$$Q_1.v_{1_{(new)}} = 0.1 * 0.80 + (1 - 0.1) * 0.89 = 0.88$$

$$Q_2.v_{2_{(new)}} = 0.1 * 0.75 + (1 - 0.1) * 0.95 = 0.93$$

Onde, $w = 1.000/10.000 = 0.1$

Depois que os valores dos critérios foram calculados, a MGQ é computada. Nesse sentido, é necessário informar o peso que cada critério exerce sobre a avaliação geral da fonte de dados. Considerando que definimos que o peso dos critérios são proporcionais temos que $w_1 = 0.5$ e $w_2 = 0.5$, a MGQ é calculada, conforme a Equação 4.7, da seguinte maneira:

$$MGQ = 0.88 * 0.5 + 0.93 * 0.5 = 0.91$$

Uma vez que os novos valores foram calculados, na Etapa 2 do processo, esses valores são atualizados em $PQ(f)$, fazendo com que as informações de qualidade da fonte de dados oferecidas aos usuários possam acompanhar a evolução que a fonte de dados sofre ao longo do tempo.

4.4 Análise Comparativa

Nesta seção, destacaremos, de maneira resumida, alguns pontos em que nossa abordagem, proposta neste capítulo, se diferencia dos trabalhos que relacionamos no Capítulo 3, em seguida sumarizamos esses pontos no Quadro 4.2.

- **Estratégia de Avaliação** - Neste trabalho estamos considerando fontes de dados com características dinâmicas e que geram um grande volume de dados ao do longo tempo. Um problema observado nesse contexto, com relação à avaliação da qualidade, é que a maioria dos trabalhos não considera esses aspectos e trata a avaliação das fontes de forma pontual.

  Levando em consideração essa problemática, principalmente considerando o aspecto dinâmico das fontes, avaliações de qualidade pontual podem não refletir o comportamento atual da fonte de dados, uma vez que elas podem sofrer modificações ao longo do tempo. Dessa forma, propomos uma avaliação da qualidade de forma contínua.

  A estratégia de avaliação contínua permite que a fonte de dados possa ser avaliada todas as vezes em que sofre uma modificação. A estratégia que estamos propondo permite que a fonte de dados seja avaliada por completo de maneira incremental e

garante que todos os dados contidos nas fontes serão avaliados, não somente uma amostra como no trabalho de Xian et al. (2009).

- **Contexto da Avaliação** - Como mostramos anteriormente, os trabalhos de Loscio et al. (2012) e Sarinho (2014) consideram requisitos da aplicação no processo de avaliação da qualidade da fonte de dados. Nesta dissertação, a avaliação da qualidade é aplicada para um contexto geral de utilização da fonte de dados, o que permite que os resultados obtidos e disponibilizados no Perfil de Qualidade, possam ser reutilizados em outros cenários.

- **Geração de Metadados de Qualidade** - Nesta dissertação propomos a geração de um conjunto de metadados, os quais chamamos de Perfil de Qualidade, com o intuito de descrever diferentes aspectos de qualidade de uma fonte de dados. Diferentemente do trabalho de Mihaila, Raschid e Vidal (2000), o qual assume que as fontes fornecem metadados semelhantes, propomos e especificamos um processo para que os metadados de qualidade possam ser gerados. Os trabalhos de Ribeiro et al. (2015) e Abele (2016) apesar de especificar um processo para geração de metadados, se diferenciam desta dissertação pois não focam especificamente na avaliação da qualidade das fontes.

Quadro 4.2: Comparação dos Trabalhos Relacionados com a Abordagem Proposta

Trabalho	Estratégia de Avaliação	Contexto da Avaliação	Geração de Metadados de Qualidade
Batista (2008)	Pontual	Específico	Não
Xian et al. (2009)	Pontual	Geral	Não
Loscio et al. (2012)	Pontual	Específico	Não
Sarinho (2014)	Mista	Específico	Não
Rekatsinas, Dong e Srivastava (2014)	Contínua	Geral	Não
Mihaila, Raschid e Vidal (2000)	Não relatado	Não relatado	Não
Ribeiro et al. (2015)	Pontual	Geral	Parcial
Abele (2016)	Pontual	Geral	Não
Este trabalho	**Contínua**	**Geral**	**Sim**

Fonte: O Autor (2016)

4.5 Considerações

Neste capítulo foi apresentada a principal contribuição desta pesquisa. Inicialmente foi mostrada uma visão geral do Perfil de Qualidade, onde o seu conceito e seus benefícios foram discutidos. Apresentamos uma categorização dos critérios de QI relevantes para avaliação de uma fonte de dados. Na sequência elencamos e especificamos os critérios completude e corretude de dados para demonstrar a estratégia de avaliação contínua. Também foi apresentada a especificação do processo para geração do Perfil de Qualidade seguido de um exemplo para ilustrar as etapas que compõe o processo.

No próximo capítulo, discutiremos a implementação de um protótipo que automatiza as etapas do processo que aqui foram propostas e apresentaremos os resultados obtidos por meio dos experimentos.

5

Implementação e Experimentos

Neste capítulo serão descritos os aspectos de implementação e experimentos realizados para validação da proposta desta dissertação. Inicialmente, descreveremos o protótipo desenvolvido, detalhando suas principais características e funcionalidades. Em seguida, apresentaremos um estudo de caso e caracterizamos as fontes de dados que serão utilizadas para realizar os experimentos. Por fim, seguiremos com uma discussão sobre os resultados obtidos.

5.1 Protótipo - *Quality Profile*

Com o objetivo de validar a estratégia de avaliação contínua e automatizar as etapas do processo de geração do Perfil de Qualidade das fontes de dados, detalhadas no capítulo anterior, desenvolvemos um protótipo, o qual chamamos de *Quality Profile*. Nesse sentido, iremos detalhar aspectos de sua implementação e funcionalidades.

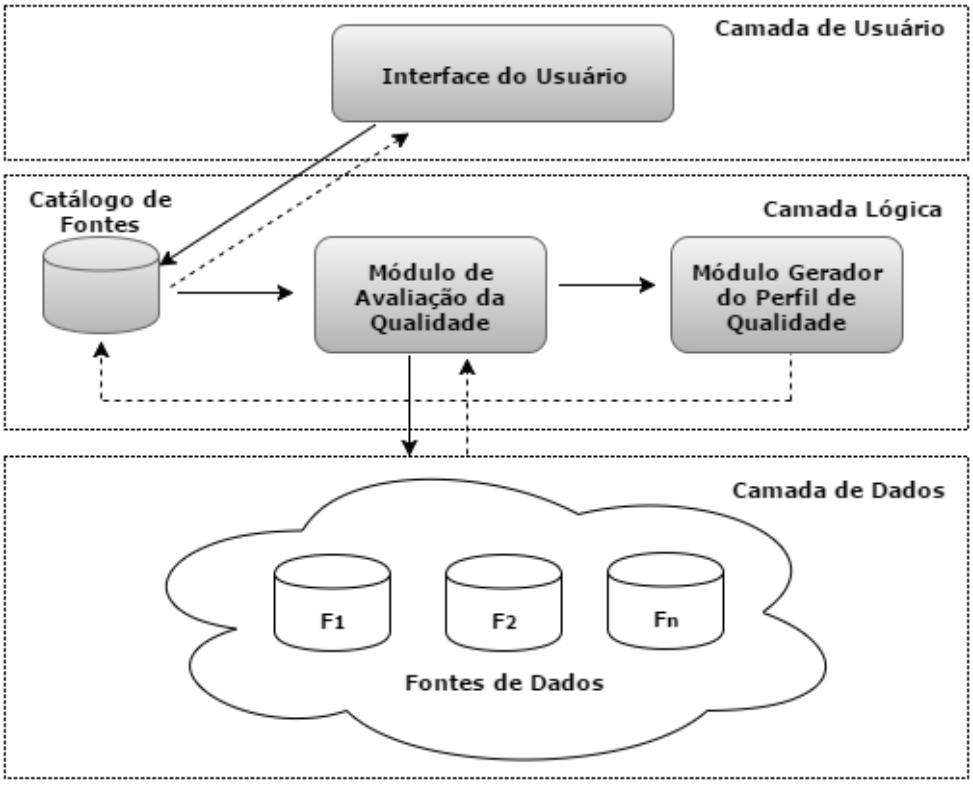
5.1.1 Desenvolvimento

O protótipo *Quality Profile* foi implementado no formato de aplicação *desktop*, por meio da linguagem JAVA[1]. A Figura 5.1 apresenta a sua arquitetura, a qual é divida em três camadas: (i) camada de usuário, responsável por fazer o usuário interagir com o sistema; (ii) camada lógica, responsável pela execução das funcionalidades do protótipo; e (iii) camada de dados, que faz a gestão das fontes de dados utilizadas no processo de avaliação da qualidade. A seguir, cada componente das camadas do protótipo será apresentado.

A **Camada de Dados** permite acesso aos dados das fontes. O acesso a esses dados pode ser feito de duas formas: remota ou local. Para que os dados possam ser acessados remotamente é necessário que a fonte de dados possua uma interface para execução de consulta ou ofereça alguma API, que será utilizada para que os dados possam ser recuperados e, posteriormente, avaliados. Entretanto, caso a fonte de dados não disponibilize tais mecanismos, os dados podem ser armazenados localmente em uma base de dados para que possam ser avaliados.

[1] https://www.java.com/

Figura 5.1: Arquitetura - *Quality Profile*



Fonte: O Autor (2016)

As fontes de dados usadas nos experimentos, realizados no protótipo, não forneciam nenhum mecanismo de acesso remoto, sendo assim, cada fonte teve seus dados armazenados em uma base de dados relacional no SGBD Oracle[2].

A **Camada Lógica** é formada por três componentes: (i) Catálogo de Fontes, (ii) Módulo de Avaliação da Qualidade, e (iii) Módulo Gerador do Perfil de Qualidade.

- **Catálogo de Fontes** - Utilizamos um catálogo de fontes para armazenar informações acerca das fontes de dados, tais como: informações estruturais (url/caminho de acesso aos dados, descrição do esquema, vocabulário, tags) e informações de qualidade, mais especificamente o Perfil de Qualidade que será gerado. Por meio dessas informações, o usuário pode analisar a fonte de dados sem necessariamente acessar diretamente os seus dados. Para armazenar as informações do catálogo de fontes utilizamos uma base de dados relacional no Oracle.

- **Módulo de Avaliação da Qualidade** - Esse componente é responsável por realizar a avaliação da qualidade nas fontes de dados. Na avaliação da qualidade foram utilizados os critérios de completude e corretude de dados e as suas respectivas métricas, detalhadas na Seção 4.2, foram implementadas em forma de *Stored Procedure* utilizando a linguagem PL/SQL[3].

[2]https://www.oracle.com/

[3]Linguagem procedural da Oracle que estende a linguagem SQL.

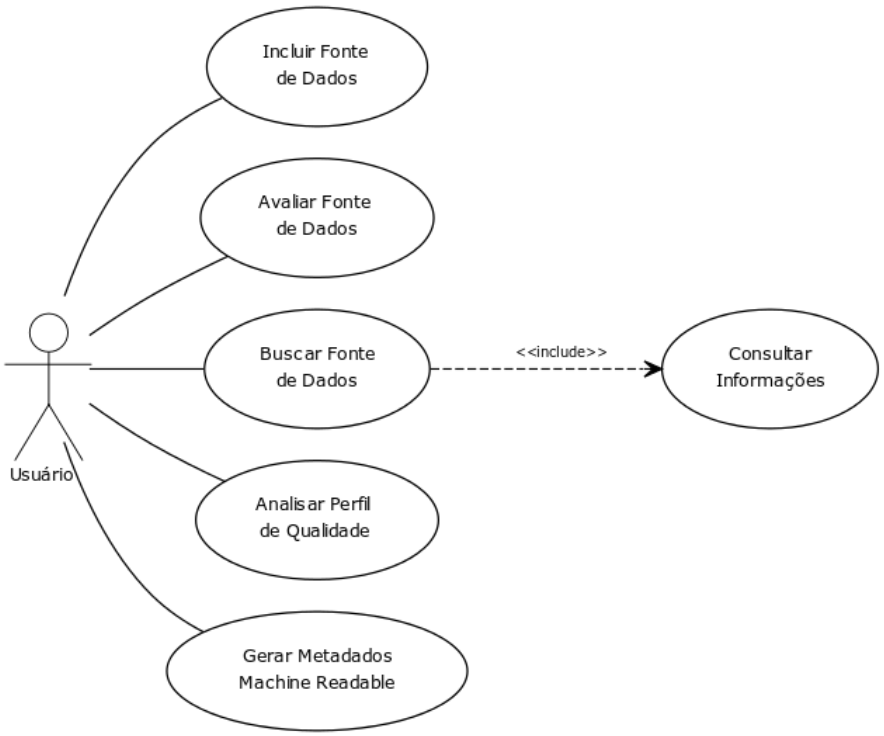
- **Módulo Gerador do Perfil de Qualidade** - Esse componente é responsável por realizar o cálculo dos critérios de qualidade e gerar a medida global de qualidade, segundo a estratégia descrita na Seção 4.3.2. Os cálculos para realizar tais ações também foram implementados em forma de *Stored Procedure* utilizando a linguagem PL/SQL.

A **Camada de Usuário** é formada pela Interface do Usuário, que é responsável pela interação do usuário com a camada lógica. Ao usuário, é permitido o acesso a diversas funcionalidades, dentre elas: (i) incluir uma fonte de dados no catálogo de fontes; (ii) avaliar a qualidade das fontes de dados; (iii) buscar fontes de dados; (iv) consultar informações de uma fonte de dados; (v) analisar Perfil de Qualidade; e (vi) Gerar Metadados *Machine Readable*. Tais funcionalidades serão detalhadas na próxima seção.

5.1.2 Funcionalidades

A seguir apresentamos o diagrama de caso de uso da *Quality Profile* (Figura 5.2), o qual corresponde às principais funcionalidades que foram desenvolvidas. Detalhes de cada uma das funcionalidades, que poderão ser executadas pelo usuário do protótipo, serão descritas a seguir.

Figura 5.2: Diagrama de Caso de Uso - *Quality Profile*



Fonte: O Autor (2016)

- **Incluir Fonte de Dados** - Para que uma fonte de dados possa ser avaliada é necessário que a mesma esteja catalogada. Durante o processo de catalogação algumas

informações são necessárias, tais como: URL ou local de acesso aos dados, descrição do esquema, descrição do vocabulário, entre outras. Após a primeira avaliação na fonte de dados, informações de qualidade vão ser incorporadas no catálogo e, posteriormente, vão sendo atualizadas à medida que novas avaliações são realizadas na fonte.

- **Avaliar Fontes de Dados** - Essa funcionalidade permite que a fonte de dados possa ser avaliada por meio dos critérios de qualidade. Nesse sentido, o Módulo de Avaliação da Qualidade é responsável por coletar, na fonte, os dados que serão avaliados e executar as *Stored Procedures* referentes às métricas de avaliação dos critérios. Em seguida, o Módulo Gerador do Perfil de Qualidade recebe o resultado da avaliação e executa a *Stored Procedure* referente à geração do Perfil de Qualidade. Ao término, as informações de qualidade geradas são associadas aos metadados da fonte de dados que foi avaliada.

- **Buscar Fontes de Dados** - Essa funcionalidade permite que um usuário possa buscar por uma ou mais fontes de dados que estejam no catálogo de fontes. A busca pode ser feita por meio de palavras-chaves ou tags (informação previamente definida no processo de catalogação). Também é dada opção para listar todas as fontes de dados catalogadas.

  A qualquer momento é possível que uma fonte de dados que esteja catalogada possa ter suas informações consultadas. Dessa forma, são apresentadas ao usuário suas informações estruturais e suas informações de qualidade mais atuais.

- **Analisar Perfil de Qualidade** - Essa funcionalidade permite que um usuário possa fazer uso das informações do Perfil de Qualidade para analisar a qualidade de uma ou mais fontes de dados. Para isso, o usuário seleciona a(s) fonte(s) de dados e o(s) atributo(s) de qualidade que deseja analisar. É apresentado ao usuário um gráfico com as informações desejadas. Para implementar essa funcionalidade foi utilizada a biblioteca JFree Chart[4].

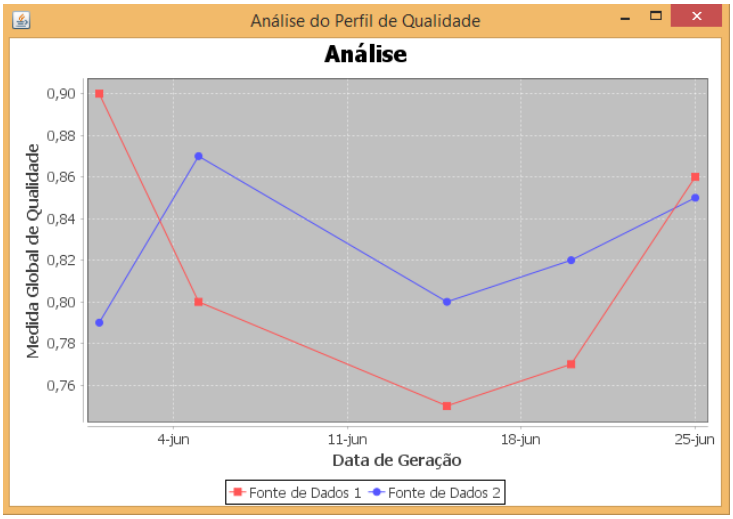
  Como a estratégia de avaliação adotada é contínua, cada vez que uma fonte de dados é avaliada suas informações de qualidade podem ser modificadas em decorrência do seu comportamento. O Perfil de Qualidade só disponibiliza o estado atual da fonte. No entanto, pode ser que informações de qualidade de momentos anteriores sejam de interesse do usuário. Dessa forma, guardamos as últimas 10 modificações do Perfil de Qualidade em um repositório auxiliar para que essas informações possam ser consultadas posteriormente.

  Nesse sentido, é dada ao usuário a opção de fazer uma análise da evolução da qualidade da fonte de dados ao longo do tempo. A Figura 5.3 mostra um exemplo de

[4]http://www.jfree.org/jfreechart/

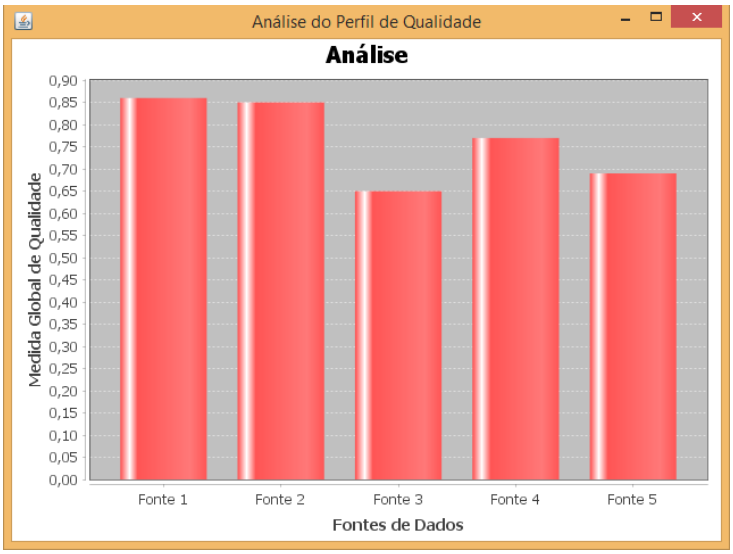
como essas informações são dispostas para o usuário. A figura em questão apresenta um comparativo da evolução de duas fontes de dados ao longo do tempo, no qual o atributo de qualidade escolhido para comparar foi a Medida Global de Qualidade. Na Figura 5.4 também é apresentado um gráfico comparativo de múltiplas fontes de dados. No entanto, considerando apenas as informações mais atuais contidas no Perfil de Qualidade das fontes de dados. Essa funcionalidade pode ser bastante útil no processo de Seleção de Fontes de Dados. Uma vez que o usuário descobre o conjunto de fontes de dados candidatas ao uso, o mesmo pode fazer uso dessa funcionalidade para compará-las.

Figura 5.3: Tela do Protótipo - Análise do Perfil de Qualidade - Evolução



Fonte: O Autor (2016)

Figura 5.4: Tela do Protótipo - Análise do Perfil de Qualidade - Em um dado momento



Fonte: O Autor (2016)

- **Gerar Metadados *Machine Readable*** - Essa funcionalidade permite que os metadados do Perfil de Qualidade possam ser gerados em formato *machine readable*. Para isso, as informações de qualidade armazenadas no catálogo são convertidas para o formato JSON, dando ao usuário a opção de copiar ou salvar o arquivo gerado.

5.2 Experimentos

Como forma de avaliar a proposta desta dissertação, mais especificamente a estratégia de avaliação contínua da qualidade das fontes de dados, realizamos alguns experimentos que serão descritos a seguir. Antes de apresentar a avaliação experimental e os resultados obtidos, apresentaremos o cenário dos experimentos e caracterizaremos as fontes de dados utilizadas.

5.2.1 Cenário e Fontes de Dados

O domínio de dados das fontes utilizadas para os experimentos foi o Meteorológico. No Brasil existem diversas agências meteorológicas que trabalham monitorando as condições climáticas em diversas áreas específicas do Brasil, por exemplo, o Instituto Nacional de Meteorologia (INMET)[5], que conta com mais de 300 estações automáticas monitorando diversas áreas do país. Outro exemplo é o Instituto de Tecnologia de Pernambuco (ITEP)[6], que monitora as condições climáticas de algumas áreas do estado de Pernambuco e a Agência Pernambucana de Águas e Clima (APAC)[7], que possui centenas de sensores monitorando toda a região do estado de Pernambuco.

As Estações Meteorológicas (EM) são responsáveis por coletar de tempos em tempos informações meteorológicas tais como: temperatura, umidade, pressão atmosférica, velocidade do vento, radiação solar, entre outras variáveis climáticas. Após a coleta, os dados são armazenados em um banco de dados e disponibilizados para consumo, que podem ser para os mais variados fins, seja para realizar previsão do tempo ou até mesmo para estudar o comportamento de uma área específica.

No entanto, nesse domínio de conhecimento, é comum que os usuários desses dados se deparem com algum tipo de problema na qualidade dos dados disponibilizados. Alguns dos principais problemas de qualidade nesse cenário ocorrem no processo de transmissão dos dados. Muitas vezes eles não são transmitidos, seja por falha dos sensores ou até mesmo o dado é transmitido, mas não reflete as condições normais, normalmente por falta de manutenção dos sensores ou falta de calibragem. Tais problemas de qualidade podem dificultar o consumo dos dados. Além disso, devido ao grande volume de dados, identificar essas falhas pode ser uma tarefa bastante complexa.

[5] http://www.inmet.gov.br/portal/
[6] http://www.itep.br/
[7] http://www.apac.pe.gov.br/meteorologia/

Neste trabalho, utilizamos duas (2) fontes de dados de duas (2) instituições que monitoram as condições climáticas da cidade do Recife (APAC e ITEP) para criar o seu respectivo Perfil de Qualidade e validar a estratégia de avaliação contínua. A escolha por essas fontes se deu justamente por elas serem dinâmicas, uma vez que atualizações de inserção são frequentes e pelo volume de dados produzido, que é crescente ao longo do tempo.

Em nossos experimentos, consideramos apenas os dados fornecidos pelos sensores de temperatura. Para avaliar a qualidade dessas fontes, utilizamos os critérios de completude e corretude (Seção 4.2).

Como vimos, a completude é dada pelas métricas de densidade e cobertura. Para avaliar a densidade buscamos por instâncias com valores nulos, já para o cálculo da cobertura foi considerada a periodicidade de medição do sensor. Por exemplo, os sensores que fornecem dados para as fontes que usamos realiza medições horárias. Dessa forma, calculamos a densidade com base nessa periodicidade.

Considerando o domínio das fontes de dados avaliadas, o conjunto de regras de validação para o critério corretude foi implementado a partir de algumas validações básicas, que compreendem três regras específicas para o domínio em questão, propostas no trabalho de Baba, Vaz e Costa (2014) e que serão detalhadas a seguir:

- **Validação de Limites** - Nessa regra é verificado se os valores a serem analisados respeitam algumas condições básicas da variável que representa, ou seja, se ela representa um valor que é fisicamente possível de ser obtido. Sob este ponto de vista, pode-se restringir valores válidos dentro de um limite de valores possíveis. Por exemplo: a umidade relativa do ar deve estar em um intervalo de dados de 0% a 100%, a temperatura deve estar em um intervalo de -90ºC a +60ºC. Qualquer valor fora deste intervalo está obviamente incorreto, pois representa um valor impossível, para a variável em questão. Nos experimentos realizados, a variável em questão foi temperatura e os limites foram definidos com base na climatologia mais recente fornecida pelo INMET para a área a qual os dados da fonte se refere, no caso Recife. Dessa forma, foram definidos os limites +10º a +40º.

- **Validação Lógica** - Muitos dos registros de uma EM registram dados médios, máximos e minimos, referentes ao intervalo do tempo de monitoramento. Estes dados devem respeitar suas condições lógicas básicas. Por exemplo, a temperatura média não pode ser maior que a temperatura máxima ou menor que a temperatura mínima desse mesmo período. Nesse caso, vale a regra $Temp_{mnima} \leq Temp_{mdia} \leq Temp_{mxima}$

- **Validação de Limites do Período** - Nessa regra busca-se identificar valores com suspeita de erros. Com base na comparação dos valores de um intervalo de tempo específico é verificado se os mesmos apresentam condições físicas realmente possível de acontecerem. Por exemplo, mesmo que as medições estejam de acordo com

as outras regras de validação, ainda assim deve-se verificar se a diferença entre a temperatura máxima e mínima, ocorrida em um mesmo período, respeita as condições físicas possíveis. Sendo assim, nessa regra foi definida, em contato com especialistas do domínio, que a diferença de temperatura máxima permitida para uma hora de medição é de 10ºC. Sendo assim, pode-se dizer que a regra $Temp_{mxima} - Temp_{mnima} \leq 10$ºC.

As instituições que disponibilizaram suas fontes de dados para os experimentos não dispunham de uma API ou um serviço que permitisse o acesso aos dados em tempo real, não sendo assim 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.

5.2.2 Avaliação Experimental

Todos os experimentos foram realizados em um computador com processador Intel Core i5 2.40GHz e 4GB de memória. 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. Nesse sentido, implementamos uma funcionalidade no protótipo *Quality Profile*, chamado Carga de Dados.

Com essa funcionalidade, quando um novo lote de dados é inserido na fonte de dados os procedimentos da avaliação de qualidade e da atualização dos metadados do Perfil de Qualidade são executados automaticamente. Para isso, implementamos uma *trigger*[8] que é responsável por realizar esse gerenciamento.

A Tabela 5.1 apresenta a quantidade de instâncias inseridas nas fontes de dados em cada instante de tempo. É importante destacar que, nesse caso, a fonte de dados seria o banco de dados que armazena os dados coletados pelo sensor e não o sensor propriamente dito.

Tabela 5.1: Lotes de Dados

Lote 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

Fonte: O Autor (2016)

[8] É um recurso de programação executado sempre que o evento associado ocorrer.

Em nossos experimentos, consideramos três (3) estratégias para realizar a avaliação da qualidade das fontes:

- **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. (2009). Especificamente, a amostra retirada para avaliação corresponde às novas instâncias incluídas na fonte após a última avaliação de qualidade da fonte;
- **Avaliação Ideal** - A estratégia que reavalia a fonte por completo todas as vezes que novas instâncias são recebidas;
- **Avaliação Contínua** - A estratégia proposta nesta dissertação, descrita no Capítulo 4.

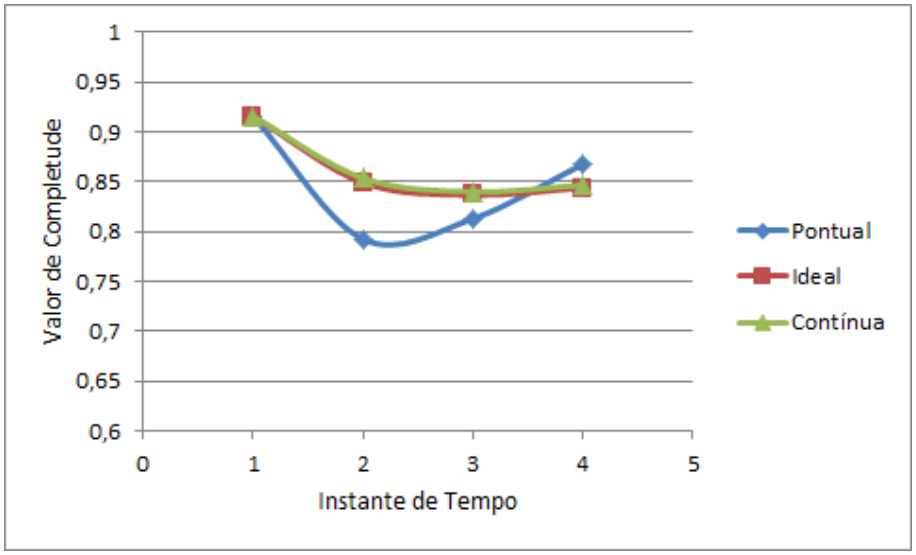
Os experimentos objetivaram validar as seguintes hipóteses:

- **H1** - A estratégia de avaliação contínua, quando comparada com a estratégia de avaliação pontual, fornece resultados mais próximos da estratégia de avaliação 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 estratégia de avaliação ideal.

Para cada instante de tempo (t_1, t_2, t_3 e t_4), realizamos a avaliação da qualidade da fonte de dados utilizando as três (3) estratégias mencionadas acima e foram obtidos os resultados ilustrados nas figuras 5.5, 5.6, 5.7 e 5.8. As figuras 5.5 e 5.6 ilustram os resultados obtidos para os critérios completude e corretude, respectivamente, para a fonte de dados APAC. As figuras 5.7 e 5.8 ilustram os resultados obtidos para os critérios Completude e Corretude, respectivamente, para a fonte de dados ITEP.

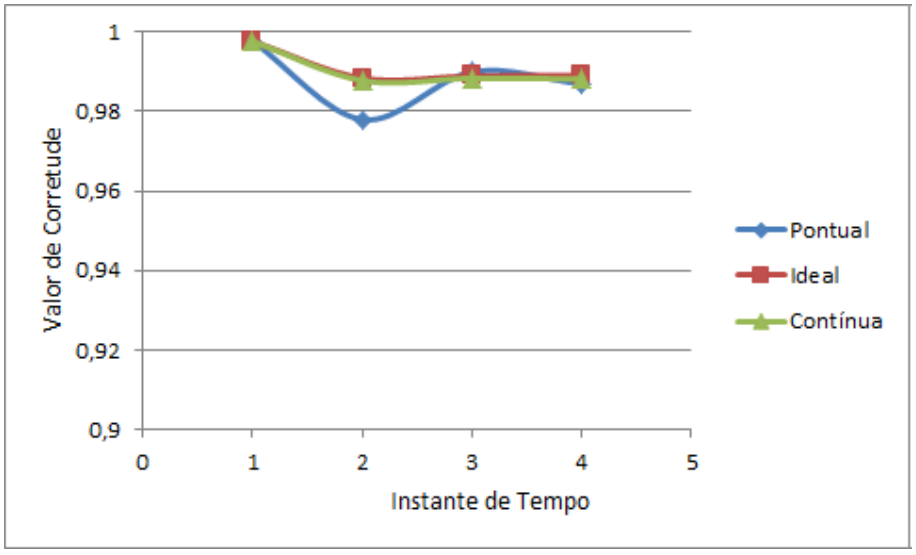
Durante a execução dos experimentos, calculamos o tempo gasto para realizar a avaliação da qualidade das fontes de dados utilizando as três (3) estratégias mencionadas anteriormente. Os tempos gastos para execução de cada estratégia em cada fonte de dados são apresentados nas figuras 5.9 e 5.10. É importante salientar que para cada instante de tempo, executamos cada estratégia cinco (5) vezes para realizar a média ponderada dos tempos obtidos em cada uma delas.

Figura 5.5: Completude - APAC



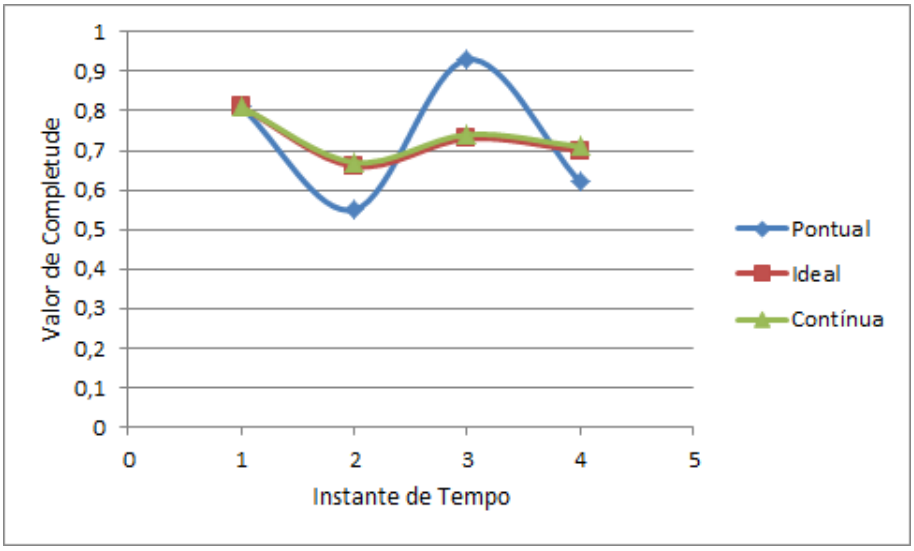
Fonte: O Autor (2016)

Figura 5.6: Corretude - APAC



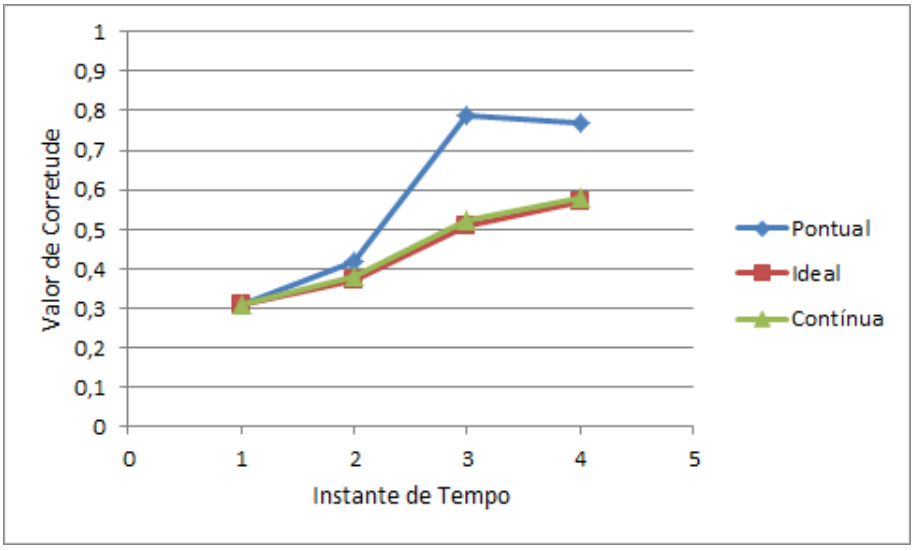
Fonte: O Autor (2016)

Figura 5.7: Completude - ITEP



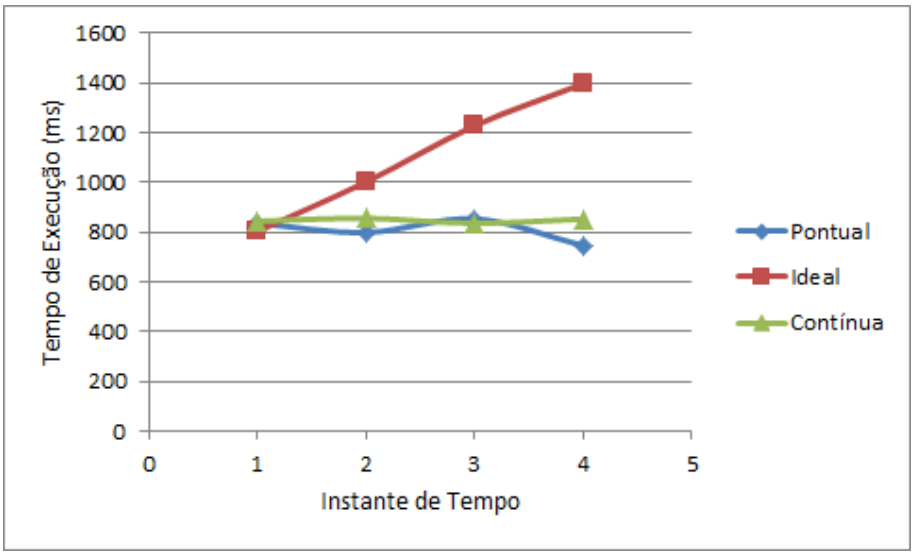
Fonte: O Autor (2016)

Figura 5.8: Corretude - ITEP



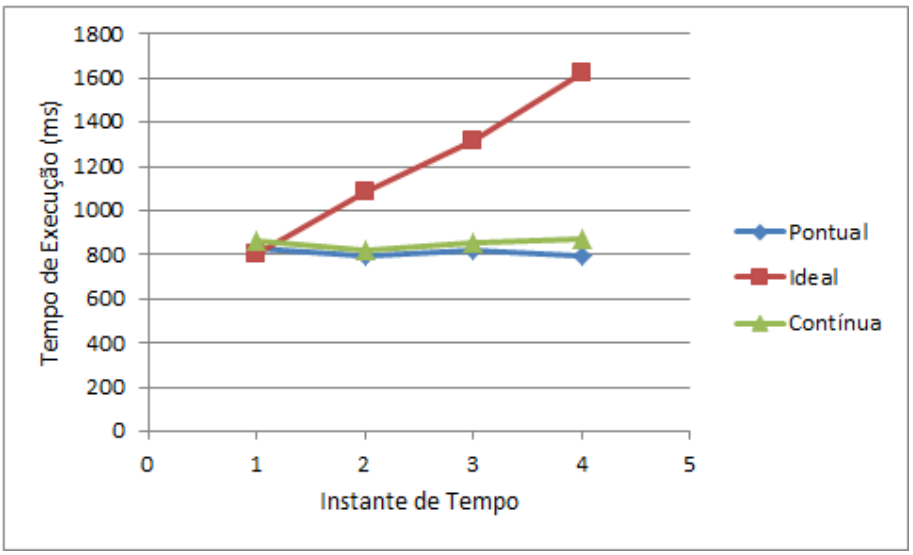
Fonte: O Autor (2016)

Figura 5.9: Tempo de Execução da Avaliação da Qualidade - APAC



Fonte: O Autor (2016)

Figura 5.10: Tempo de Execução da Avaliação da Qualidade - ITEP



Fonte: O Autor (2016)

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 necessidade de uma avaliação contínua da qualidade em fonte 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 (figuras 5.5 e 5.6), quanto na do ITEP

(figuras 5.7 e 5.8), os valores atribuídos aos critérios de qualidade, em diferentes instantes de tempo, tiveram valores distantes do esperado quando comparados com a estratégia ideal.

Por exemplo, na avaliação da fonte de dados ITEP para o critério completude, no instante de tempo t_3 (Figura 5.7), o valor correto do critério seria 0.73 enquanto a estratégia de avaliação pontual calculou o valor 0.93, atribuindo um valor 23,39% maior do que realmente era para ser. 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 (figuras 5.5, 5.6, 5.7 e 5.8). Em todos os instantes de tempo, os valores atribuídos aos critérios de qualidade tiveram valores bem próximos do esperado, quando comparados com a estratégia ideal.

Tomando como ilustração o exemplo anterior, no mesmo instante de tempo o valor atribuído ao critério pela estratégia de avaliação contínua foi 0.74, sendo esse valor 1,36% maior do que realmente era pra ser, uma diferença muito baixa. 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 essas análises pudemos validar a hipótese H1. Verificando os gráficos gerados com os valores dos critérios de qualidade é possível observar que a diferença entre os valores da estratégia de avaliação contínua com relação à estratégia de avaliação ideal é muito baixa, tendo variações de +/- 1 a 2%, o que caracteriza que a precisão dos valores obtidos pela estratégia proposta nesta dissertação é alta, validando assim a hipótese H2.

Com relação ao custo computacional, das três estratégias, a avaliação pontual apresentou um menor custo. 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 de tempo t_4 houve um aumento de 12% (APAC) e 9% (ITEP) (figuras 5.9 e 5.10).

Analisando os resultados apresentados nas figuras 5.9 e 5.10, é possível observar que a estratégia de 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 considéravel no tempo de execução da avaliação.

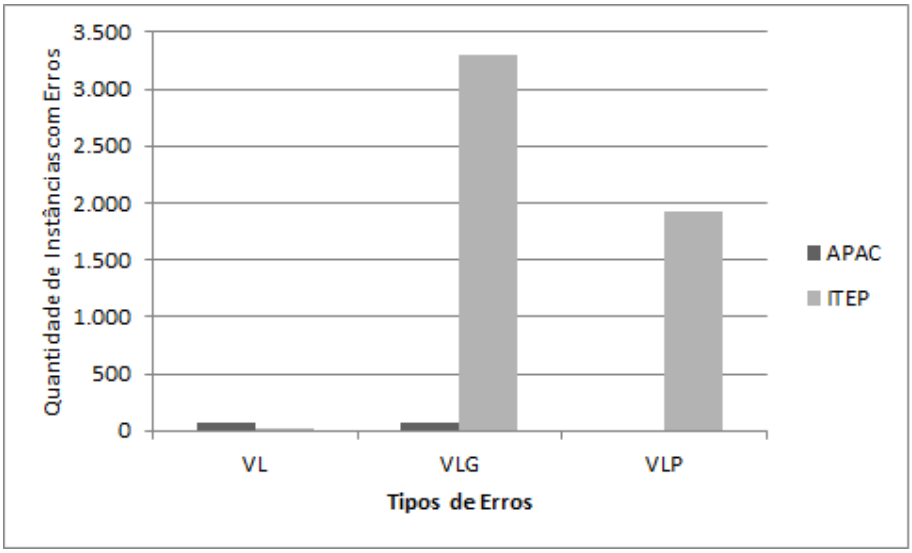
Por exemplo, quando comparamos o instante t_4, na qual o volume de dados acumulado nas fontes é maior, há uma redução no tempo de execução da avaliação de 63% (APAC) e 86% (ITEP) sem que haja uma perda de precisão considerável nos valores dos critérios. Com essa análise foi possível validar a hipótese H3. Acreditamos que esse ganho pode ainda ser maior se a

avaliação considerar um conjunto grande de fontes e um grande volume de dados.

É importante também apresentar uma visão geral do que foi possível observar na qualidade dos dados das fontes utilizadas. Com relação à avaliação dos critérios de qualidade, no critério completude foi possível verificar, em ambas as fontes de dados, que os problemas eram mais recorrentes na métrica de cobertura do que na densidade. Isso se deve ao fato de falhas que ocorreram nos sensores no processo de transmissão das medições. Esse tipo de situação acontece devido à indisponibilidade de rede móvel nos sensores, fazendo com que os dados não sejam transmitidos para a base de dados e a informação se torne incompleta para uma análise posterior dos dados.

O critério corretude apresentou variações nos resultados. A Figura 5.11 apresenta um gráfico em barras com a quantidade de instâncias com erros encontradas durante a avaliação geral das fontes. Agrupamos os valores por tipos de erro, onde VL corresponde à restrição de Validação de Limites, VLG à Validação Lógica e VLP à Validação de Limites do Período. As regras foram detalhadas na Seção 5.2.1.

Figura 5.11: Tipos de Erros nas Instâncias das Fontes de Dados



Fonte: O Autor (2016)

Na fonte de dados APAC, a quantidade de instâncias com mais erros foi encontrada por meio da regra da Validação de Limites. Isso ocorreu porque muitas vezes o sensor atribuiu o valor de temperatura mínima com valores muito baixos. Há medições com valores registrados com -1º e 5º, por exemplo. Com relação à fonte de dados ITEP, os erros mais recorrentes foram encontrados por meio da Validação Lógica. Isso ocorreu porque em vários períodos o sensor atribuiu valores à temperatura média maior que a temperatura máxima, ocasionando assim um erro quando comparava os respectivos valores.

5.4 Considerações

Neste capítulo discutimos a implementação da estratégia proposta e os experimentos realizados para a sua avaliação. Inicialmente apresentamos o protótipo *Quality Profile*, onde detalhamos a sua arquitetura e suas principais funcionalidades. Logo em seguida, a avaliação experimental foi descrita, na qual caracterizamos o cenário e as fontes de dados utilizadas, e elencamos as hipóteses a serem validadas.

Os 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. No próximo capítulo apresentaremos as conclusões e indicações de trabalhos futuros.

6

Conclusão

Neste capítulo realizamos as considerações finais sobre esta dissertação. Apresentamos uma visão geral do trabalho, as contribuições de pesquisas alcançadas e indicações de trabalhos futuros.

6.1 Visão Geral

Identificar e selecionar fontes de dados para um determinado uso tornou-se um desafio, devido ao grande número de fontes de dados disponíveis. Nesse sentido, a literatura oferece diversos trabalhos que propõem abordagens que utilizam critérios de QI como meio para avaliar a qualidade das fontes de dados, e com base nisso, identificar aquela(s) que é/são mais adequada(s) a uma determinada necessidade.

Ao longo da pesquisa, apresentamos um levantamento dos principais trabalhos oferecidos na literatura que realizam avaliações de qualidade nas fontes de dados e identificamos que a grande maioria faz uso de estratégias de avaliações pontuais, ou seja, só consideram um instante de tempo específico e/ou não consideram o aspecto dinâmico das fontes de dados.

Dessa forma, nesta dissertação focamos no problema de avaliação da qualidade de fontes dinâmicas e vimos que estratégias de avaliações pontuais não são suficientes para representar com precisão a qualidade geral de tais fontes de dados. Como solução, propomos uma estratégia de avaliação contínua. A solução apresentada leva em consideração o estado atual da fonte de dados e resultados de avaliações que ocorreram em momentos anteriores.

Vimos também que informações de qualidade das fontes de dados podem ser úteis para diversos fins, inclusive para o processo de seleção de fontes. Nesse sentido, também propomos o Perfil de Qualidade. Conforme apresentamos, o Perfil de Qualidade é definido como um conjunto de metadados que descrevem a qualidade de uma fonte de dados. Apesar de seu uso servir para múltiplas tarefas, no levantamento realizado na literatura não encontramos trabalhos que propõem um processo para geração desse tipo específico de metadados.

Formalizamos o conceito e especificamos um processo para geração do Perfil de Qualidade. O processo, que foi dividido em duas etapas, faz uso da estratégia de avaliação contínua,

proposta nesta dissertação. Apresentamos também um protótipo, denominado *Quality Profile*, cujo objetivo é automatizar as etapas do processo especificado.

Utilizamos fontes de dados do domínio de Meteorologia para validar as hipóteses iniciais, bem como a estratégia de avaliação contínua. Por meio dos experimentos realizados foi possível confirmar que ao longo do tempo a qualidade das fontes de dados mudam em decorrência das modificações que ocorrem em seu comportamento. Utilizamos três estratégias de avaliação e concluímos que no cenário apresentado a estratégia de avaliação contínua se mostrou mais eficiente, uma vez que conseguiu apresentar *trade-off*, entre precisão e custo, melhores quando comparada com as demais estratégias avaliadas. Assim, conclui-se que o presente trabalho conseguiu atingir seus objetivos.

6.2 Contribuições Alcançadas

Dentre as principais contribuições deste trabalho, destacam-se:

- **Estratégia de Avaliação Contínua** - Como vimos, o custo para revaliar a fonte de dados por completo todas as vezes que a mesma se modifica pode ser alto, principalmente quando a fonte gera um grande volume de dados ao longo do tempo. Nesse sentido, propomos uma estratégia de avaliação contínua, que segue um viés incremental. À medida que a fonte de dados vai sendo modificada, as novas instâncias são avaliadas e os resultados de avaliações anteriores são utilizados para atribuir um valor de qualidade para a fonte, sem ter que reavaliá-la por completo todas as vezes que a mesma é modificada.

- **Perfil de Qualidade** - Por meio de um levantamento realizado na literatura, foi possível identificar uma grande quantidade de trabalhos que utilizam informações de qualidade das fontes de dados para diversos fins. Apesar da W3C[1] indicar como uma boa prática que ao ser publicada uma fonte de dados informações que descrevam a sua qualidade também seja disponibilizada (LOSCIO; BURLE; CALEGARI, 2016), a grande maioria das fontes encontradas não possui tais informações. Dessa forma, especificamos um processo para que um Perfil de Qualidade possa ser gerado para fontes de dados. O Perfil de Qualidade reúne informações que descrevem a qualidade de uma fonte e pode ser disponibilizado juntamente com a fonte de dados, fazendo com que essas informações disponibilizadas possam ser utilizadas pelos consumidores de dados de múltiplas formas.

- **Protótipo *Quality Profile*** - Foi implementado um protótipo para automatizar o processo de geração do Perfil de Qualidade. Também foi apresentado algumas

[1] https://www.w3.org/

funcionalidades, dentre elas a de Análise do Perfil de Qualidade, que faz uso das informações de qualidade para comparar um conjunto de fontes de dados.

É importante destacar que, a partir dos resultados obtidos durante o desenvolvimento desta dissertação, até a presente data, foi gerada a seguinte publicação para a comunidade científica:

- SILVA NETO, E. C.; Lóscio, B. F.; Salgado, A. C. **Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas**. In: *Proceedings of 31º Simpósio Brasileiro de Banco de Dados - SBBD*. Salvador - BA, Brasil, 2016. (Aceito como *full paper* para publicação em Outubro/2016)

6.3 Limitações e Trabalhos Futuros

Nesta seção destacamos algumas limitações do presente trabalho. Nesse sentido, apresentamos algumas direções que podem ser exploradas em trabalhos futuros.

- **Considerar outras características de modificações** - Neste trabalho só tratamos como modificações a inclusão de dados nas fontes. Nesse sentido, pretendemos ampliar o escopo e considerar também ações de atualização e remoção de dados.

- **Expansão dos Critérios de QI** - Neste trabalho avaliamos apenas a completude e a corretude das fontes de dados. Para enriquecer as informações contidas no Perfil de Qualidade, de forma a proporcionar uma avaliação de qualidade mais ampla, pretende-se expandir o conjunto de critérios de qualidade, incorporando critérios relacionados à qualidade do serviço, tais como: disponibilidade, tempo de resposta e desempenho (ZHU; BUCHMANN, 2002; DUSTDAR et al., 2012).

- **Realizar novos experimentos** - Para validar nossa proposta utilizamos apenas fontes de dados do domínio Meteorológico. Pretendemos ampliar os experimentos para outros domínios de dados. Um cenário de experimento preliminar já começou a ser estudado. Nesse experimento utilizaremos dados de GPS dos ônibus da cidade do Rio de Janeiro, fornecidos pelo portal de dados abertos do Rio[2]. Nesse novo experimento, pretende-se avaliar outros critérios de QI, como o frescor das fontes de dados (REKATSINAS; DONG; SRIVASTAVA, 2014).

- **Avaliar a usabilidade do Perfil de Qualidade** - As informações do Perfil de Qualidade podem ser exploradas para diversos fins, como (i) no processo de seleção de fontes de dados (XIAN et al., 2009; LOSCIO et al., 2012; DONG; SAHA; SRIVASTAVA, 2012); e (ii) no enriquecimento de consulta com atributos de qualidade

[2]http://data.rio/dataset

(YEGANEH et al., 2009). Nesse sentido, a proposta é montar cenários nos contextos citados para realizar experimentos com o objetivo de avaliar a usabilidade do Perfil de Qualidade para tais fins.

- **Expansão da ferramenta *Quality Profile*** - Incluir a solução proposta nesta dissertação em uma ferramenta maior. A ideia é montar um catálogo de fontes de dados, semelhante ao proposto em Oliveira, Gama e Loscio (2015), na qual o catálogo possa ser utilizado em soluções de Integração de Dados e ofereça aos seus usuários um conjunto de metadados sobre as fontes, inclusive metadados de qualidade (Perfil de Qualidade) que serão gerados pela estratégia proposta nesta dissertação.

Referências

ABEDJAN, Z.; GOLAB, L.; NAUMANN, F. Profiling relational data: A survey. *The VLDB Journal*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 24, n. 4, p. 557–581, 2015. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/s00778-015-0389-y>.

ABELE, A. Linked data profiling: Identifying the domain of datasets based on data content and metadata. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016. p. 287–291. Disponível em: <http://dx.doi.org/10.1145/2872518.2888603>.

AKSOY, D. Information source selection for resource constrained environments. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 34, n. 4, p. 15–20, 2005. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/1107499.1107500>.

ARAZY, O.; KOPAK, R. On the measurability of information quality. *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, v. 62, n. 1, p. 89–99, 2011. ISSN 1532-2890. Disponível em: <http://dx.doi.org/10.1002/asi.21447>.

BABA, R. K.; VAZ, M. S. M. G.; COSTA, J. Correção de dados agrometeorológicos utilizando métodos estatísticos. *Revista Brasileira de Meteorologia*, v. 29, n. 4, 2014.

BASTOS, M. R. et al. Data integration: Quality aspects. In: *Transmission and Distribution Conference and Exposition: Latin America (T D-LA), 2010 IEEE/PES*. São Paulo - SP - Brazil: IEEE, 2010. p. 411–416.

BATINI, C. et al. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 41, n. 3, p. 16:1–16:52, 2009. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1541880.1541883>.

BATISTA, M. d. C. M. *Schema Quality Analysis in Data Integration System*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2008.

BHARAMBE, D.; JAIN, S.; JAIN, A. A survey: Detection of duplicate record. *International Journal of Emerging Technology and Advanced Engineering*, v. 2, p. 298–307, 2012. ISSN 2250-2459.

BOUZEGHOUB, M.; PERALTA, V. A framework for analysis of data freshness. In: *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*. New York, NY, USA: ACM, 2004. (IQIS'04), p. 59–67. Disponível em: <http://doi.acm.org/10.1145/1012453.1012464>.

CAI, L.; ZHU, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, v. 14, p. 2, 2015. ISSN 1683-1470. Disponível em: <http://datascience.codata.org/articles/10.5334/dsj-2015-002/>.

DASU, T. et al. Mining database structure; or, how to build a data quality browser. In: *Proceedings of the 2002 ACM SIGMOD International Conference on Management of*

Data. New York, NY, USA: ACM, 2002. (SIGMOD'02), p. 240–251. Disponível em: <http://doi.acm.org/10.1145/564691.564719>.

DEBATTISTA, J.; LANGE, C.; AUER, S. daq, an ontology for dataset quality information. In: *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)*. Seoul, Korea: ACM, 2014. Disponível em: <http://ceur-ws.org/Vol-1184/ldow2014_paper_09.pdf>.

DONG, X. L.; NAUMANN, F. Data fusion: Resolving data conflicts for integration. *Proc. VLDB Endow.*, VLDB Endowment, Lyon, France, v. 2, n. 2, p. 1654–1655, 2009. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.14778/1687553.1687620>.

DONG, X. L.; SAHA, B.; SRIVASTAVA, D. Less is more: Selecting sources wisely for integration. *Proc. VLDB Endow.*, VLDB Endowment, Instabul, v. 6, n. 2, p. 37–48, 2012. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.14778/2535568.2448938>.

DONG, X. L.; SRIVASTAVA, D. *Big Data Integration*. [S.l.]: Morgan & Claypool, 2015. v. 1.

DUQUENNOY, S.; GRIMAUD, G.; VANDEWALLE, J. J. The web of things: Interconnecting devices with high usability and performance. In: *Embedded Software and Systems, 2009. ICESS '09. International Conference on*. Hangzhou, Zhejiang, China: [s.n.], 2009. p. 323–330.

DUSTDAR, S. et al. Quality-aware service-oriented data integration: Requirements, state of the art and open challenges. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 41, n. 1, p. 11–19, 2012. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/2206869.2206873>.

HELLERSTEIN, J. M. et al. The madlib analytics library: Or mad skills, the sql. *Proc. VLDB Endow.*, VLDB Endowment, v. 5, n. 12, p. 1700–1711, 2012. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.14778/2367502.2367510>.

JOHNSON, T. *Encyclopedia of Database Systems*. 2009. Chapter Data Profiling.

LOSCIO, B. F. et al. 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*. Bali, Indonesia: ACM, 2012. (IIWAS'12), p. 36–44. ISBN 978-1-4503-1306-3. Disponível em: <http://doi.acm.org/10.1145/2428736.2428747>.

LOSCIO, B. F.; BURLE, C.; CALEGARI, N. *Data on the Web Best Practices*. 2016. Disponível em: <https://www.w3.org/TR/dwbp/>.

MALAVERRI, J. E. G.; SANTANCHE, A.; MEDEIROS, C. B. A provenance-based approach to evaluate data quality in escience. *Int. J. Metadata Semant. Ontologies*, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 9, n. 1, p. 15–28, 2014. ISSN 1744-2621. Disponível em: <http://dx.doi.org/10.1504/IJMSO.2014.059127>.

MIHAILA, G. A.; RASCHID, L.; VIDAL, M.-E. Using quality of data metadata for source selection and ranking. In: *Vossen (Eds.), Proceedings of the Third International Workshop on the Web and Databases, WebDB*. Dallas, Texas, USA: [s.n.], 2000. p. 93–98.

NAUMANN, F. Data fusion and data quality. In: *Proceedings of the New Techniques and Technologies for Statistics Seminar (NTTS'98)*. Sorrent, Italy: [s.n.], 1998.

NAUMANN, F. Data profiling revisited. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 42, n. 4, p. 40–49, 2014. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/2590989.2590995>.

NAUMANN, F.; FREYTAG, J. C. Completeness of information sources. *Technical Report HUB-IB-135*, Humboldt University of Berlin, 2000.

NAUMANN, F.; FREYTAG, J. C.; SPILIOPOULOU, M. Quality-driven source selection using data envelopment analysis. In: *Proceedings of the 3rd Conference on Information Quality (IQ)*. [S.l.: s.n.], 1998. p. 137–152.

NAUMANN, F.; ROLKER, C. Assessment methods for information quality criteria. In: *Proceedings of the 5th Conference on International Quality (IQ)*. Boston, USA: [s.n.], 2000. p. 148–162.

OLIVEIRA, H. R. d. *Explorando o Feedback do Usuário para Classificação de Fontes de Dados em Sistemas de Integração Pay-as-you-go*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2012.

OLIVEIRA, M. I. S.; GAMA, K. S. da; LOSCIO, B. F. Waldo: Data producers registry and discovery service for smart cities middleware. In: *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems*. Goiania, Goias, Brasil: Brazilian Computer Society, 2015. (SBSI 2015), p. 71–78. Disponível em: <http://dl.acm.org/citation.cfm?id=2814058.2814071>.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. *Commun. ACM*, ACM, New York, NY, USA, v. 45, n. 4, p. 211–218, 2002. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/505248.506010>.

PIRES, C. E. et al. *Data Management in Grid and Peer-to-Peer Systems: Second International Conference*. Berlin, Heidelberg: Springer, 2009. 124–135 p. ISBN 978-3-642-03715-3. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03715-3_11>.

RAMAN, V.; HELLERSTEIN, J. M. Potter's wheel: An interactive data cleaning system. In: *Proceedings of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (VLDB'01), p. 381–390. ISBN 1-55860-804-4. Disponível em: <http://dl.acm.org/citation.cfm?id=645927.672045>.

REKATSINAS, T. et al. Finding quality in quantity: The challenge of discovering valuable sources for integration. In: *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research*. Asilomar, California: ACM, 2015. Disponível em: <http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper21.pdf>.

REKATSINAS, T.; DONG, X. L.; SRIVASTAVA, D. Characterizing and selecting fresh data sources. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. Snowbird, Utah, USA: ACM, 2014. (SIGMOD'14), p. 919–930. ISBN 978-1-4503-2376-5. Disponível em: <http://doi.acm.org/10.1145/2588555.2610504>.

RIBEIRO, D. C. et al. An urban data profiler. In: *Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy: ACM, 2015. (WWW '15 Companion), p. 1389–1394. ISBN 978-1-4503-3473-0. Disponível em: <http://doi.acm.org/10.1145/2740908.2742135>.

SAMIR, K.; HABIBA, D. Mutli-agent system for personalizing information source selection. In: *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences on*. Milano, Italy: ACM, 2009. v. 1, p. 588–595.

SARINHO, W. T. *Uma Abordagem para Avaliação da Qualidade de Linked Datasets para Aplicações de Domínio Específico*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2014.

SIDI, F. et al. Data quality: A survey of data quality dimensions. In: *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*. Kuala Lumpur, Malaysia: IEEE, 2012. p. 300–304.

TALUKDAR, P. P.; IVES, Z. G.; PEREIRA, F. Automatically incorporating new sources in keyword search-based data integration. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. Indianapolis, Indiana, USA: ACM, 2010. (SIGMOD'10), p. 387–398. Disponível em: <http://doi.acm.org/10.1145/1807167.1807211>.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, ACM, New York, NY, USA, v. 39, n. 11, p. 86–95, 1996. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/240455.240479>.

WANG, R. Y. A product perspective on total data quality management. *Commun. ACM*, ACM, New York, NY, USA, v. 41, n. 2, p. 58–65, 1998. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/269012.269022>.

WANG, R. Y.; KON, H. B.; MADNICK, S. E. Data quality requirements analysis and modeling. In: *Proceedings of the Ninth International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1993. p. 670–677. Disponível em: <http://dl.acm.org/citation.cfm?id=645478.654796>.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, M. E. Sharpe, Inc., Armonk, NY, USA, v. 12, n. 4, p. 5–33, mar. 1996. ISSN 0742-1222. Disponível em: <http://dx.doi.org/10.1080/07421222.1996.11518099>.

WANG, X.; DONG, X. L.; MELIOU, A. Data x-ray: A diagnostic tool for data errors. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne, Victoria, Australia: ACM, 2015. (SIGMOD'15), p. 1231–1245. Disponível em: <http://doi.acm.org/10.1145/2723372.2750549>.

XIAN, X.-F. et al. Quality-based data source selection for web-scale deep web data integration. In: *2009 International Conference on Machine Learning and Cybernetics*. [S.l.: s.n.], 2009. v. 1, p. 427–432. ISSN 2160-133X.

YEGANEH, N. K. et al. Data quality aware queries in collaborative information systems. In: *Advances in Data and Web Management: Joint International Conferences, APWeb/WAIM 2009*. Suzhou, China: [s.n.], 2009.

ZAVERI, A. et al. Quality assessment for linked data: A survey. *Semantic Web Journal*, 2012. Disponível em: <http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey>.

ZHU, Y.; BUCHMANN, A. Evaluating and selecting web sources as external information resources of a data warehouse. In: *Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on*. [S.l.: s.n.], 2002. p. 149–160.