Automatização da Extração de Dados na Open Smart City View Utilizando Crawling

Fernanda Rigo, Roberto S. Rabello, Ericles A. Bellei, Fábio L. Brezolin

Programa de Pós-Graduação em Computação Aplicada (PPGCA)
Universidade de Passo Fundo – Caixa Postal 611 – 99.001-970 – Passo Fundo – RS
{64379, rabello, 168729, 71856}@upf.br

Abstract. *In the context of smart cities, the citizen is an important agent amidst of the information and knowledge provided by open government data. Open Smart City View (OSCV) is a technological model for collecting, processing and presenting information from governmental portals. To complement the functionalities of OSCV, this work presents a crawler-based way to automate the data collection for presenting results in an accessible interface to end-user, providing a better citizen engagement by making it more participatory in society.*

Resumo. *No contexto das cidades inteligentes, o cidadão mostra-se um importante agente em meio às informações e conhecimentos disponibilizados pelos dados governamentais abertos. Nesse cenário, surge arquitetura Open Smart City View (OSCV) como um modelo tecnológico de coleta, processamento e apresentação de informações oriundas de portais governamentais. Para complementar as funcionalidades da OSCV, propõe-se um método baseado em crawling para automatizar a coleta de dados na apresentação do resultado ao usuário final em uma interface de fácil interpretação, propiciando, assim, um melhor engajamento do cidadão ao torná-lo mais participativo na sociedade.*

1. Introdução

A participação e o engajamento dos cidadãos nos assuntos relacionados a governança são fundamentais para a evolução das cidades ao torná-las cada vez mais inteligentes. Muitas são as informações que se encontram em poder dos órgãos governamentais e que necessitam estar disponíveis aos cidadãos para que, assim, possam usufruir de seu conteúdo. Os portais de transparência são os meios utilizados para disponibilizar as informações de ordem pública, mas que ainda deixam a desejar em relação à estruturação do conteúdo, dificultando o entendimento e mensuração dos dados.

Com objetivo de auxiliar o cidadão a usufruir dos dados disponibilizados pelos poderes públicos de forma mais clara e interativa, elaborou-se a arquitetura OSCV (Open Smart City View) [Lusa 2016]. A arquitetura foi desenvolvida e continua sendo pesquisada dentro do Programa de Pós-Graduação em Computação Aplicada na Universidade de Passo Fundo. Seu objetivo é viabilizar ao cidadão conhecer e avaliar os assuntos de seu município, por meio de apresentações simplificadas das informações coletadas de dados governamentais abertos. Essa arquitetura é um modelo tecnológico funcional de coleta, processamento e apresentação de dados em uma interface *web*.

A extração de dados é um fator indispensável para o processo funcional da arquitetura OSCV, pois além de realizar a coleta, é responsável por manter os dados em um *data warehouse* atualizado. Nesse sentido, este artigo tem por objetivo demonstrar as

técnicas que estão sendo utilizadas para integrar a evolução da arquitetura OSCV, com a extração automatizada na coleta de informações a partir dos arquivos disponibilizados pelos portais de transparência.

2. Fundamentação Teórica e Definições Conceituais

A cidade tem sido a opção de um grande número de pessoas que viviam no campo, pois no meio urbano é maior a variedade de emprego, facilidade de acesso à saúde, educação, entretenimento e cultura [Berst 2013]. Com o aumento da população, os centros urbanos precisam estar em constante aprimoramento e evolução para atender às demandas de seus habitantes. Os governos enfrentam diversos desafios nesse crescimento, muitas vezes desordenado. Com isso, torna-se essencial para a construção de uma cidade competitiva o desenvolvimento de setores como infraestrutura, tecnologias, serviços, governo, recursos naturais, entre outros [Avelar 2013].

[Gray 2010] afirma que a participação do cidadão é fundamental para o processo de inovações tecnológicas, sociais, culturais, econômicas. As tecnologias e o potencial das informações em poder governamental são ferramentas que podem auxiliar e orientar indivíduos, empresas e indústrias no desenvolvimento das cidades. A abertura dos dados governamentais à população em geral beneficia e auxilia no empoderamento do cidadão para a construção de cidades mais inteligentes e participativas. O volume de informação em poder dos órgãos públicos é elevado e muitas vezes os dados não estão minerados, o que dificulta o acesso e a compreensão aos mesmos, que se tornam, assim, inacessíveis à sociedade [Schuurman *et al.* 2012].

O cidadão inteligente é reconhecido por sua capacidade de conectar-se a cidade, interagindo por meio dos diversos canais de comunicação disponibilizados. Um dos objetivos da arquitetura Open Smart City View (OSCV) é se tornar uma importante ferramenta de empoderamento do cidadão. A OSCV é um modelo tecnológico funcional de coleta, processamento e apresentação de informações relevantes acerca de um determinado município. Os dados de entrada da OSCV são dados governamentais abertos, coletados de diversas fontes [Lusa 2016]. A arquitetura OSCV divide-se em quatro camadas funcionais de operação: Fontes de Dados, Coleta e Processamento, Web e Dispositivos de Interação (Figura 1).

A camada Fonte de Dados tem a função de identificar e mapear as fontes de dados governamentais abertos de interesse. Realiza-se uma análise prévia dos conjuntos de dados que melhor atendem as necessidades de informações desejadas. Na camada de Coleta e Processamento, os conjuntos de dados identificados na camada anterior são coletados de seus locais de origem para passar por um *pipeline* de operações com extração, transformação e carga dos registros no *data warehouse* da arquitetura, realizando um processo conhecido como ETL – *Extract, Transform and Load*. Os scripts ETL realizam a leitura dos dados em seu estado bruto, executam atividades de saneamento e transformação, e por fim, armazenam os dados resultantes nas tabelas de interface, mantendo o *data warehouse* atualizado. A camada Web tem a função de fornecer serviços de acesso e consumo para os dados armazenados, lhe conferindo características de um *software* OLAP – *Online Analytical Processing*.

Essa camada situa-se como uma interface analítica entre os dados mantidos no *data warehouse* e os diversos dispositivos de interação (camada externa) que os cidadãos utilizam para obter a informação. No tópico seguinte será apresentada a metodologia utilizada para o processo de extração de dados.

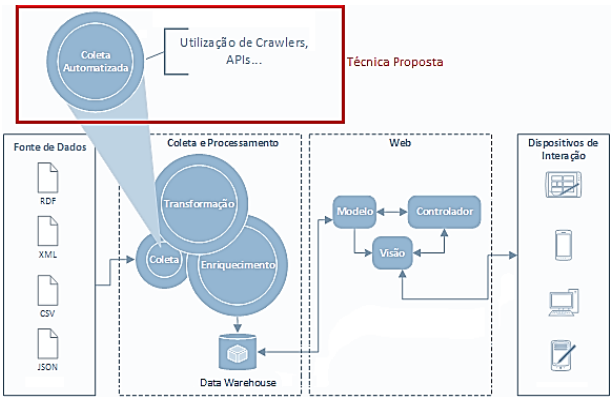

Figura 4. Modelo simplificado das camadas funcionais da arquitetura OSCV

3. A Metodologia Proposta

O objetivo da metodologia proposta é colaborar com a evolução da arquitetura OSCV no processo de coleta dos dados. Tal processo é uma atividade essencial para a arquitetura OSCV, pois seus resultados impactam diretamente na qualidade do resultado dos processos subsequentes. Os arquivos que podem ser coletados precisam estar em formatos não proprietários, como HTML, TXT, XML, JSON, CSV e RDF.

Após identificados os portais de transparência que serão utilizados como fonte para coleta dos dados desejados e implementou-se um processo de requisição e extração baseado em *crawling*. *Web crawlers* são ferramentas consideravelmente referenciadas no setor de automação e indexação de dados, pois realizam tarefas de agentes de *softwares* utilizados para especializar pesquisas, capturas e extração de dados.

O *web crawler* foi implementado na linguagem Java, mais especificamente com a API HtmlUnit. Essa biblioteca tem funcionalidades para automatizar ações de navegador, além de permitir a execução de comandos e ações nos endereços desejados dos portais, da mesma forma que um usuário comum realiza a navegação por browser. Realiza-se consultas em um navegador invocando páginas *web*, varreduras de dados em elementos HTML, cliques em botões, links entre outras ações.

O processo de extração automatizada contempla a configuração das várias fontes a serem extraídas e a sequência de comandos a serem executados para a correta busca e extração de cada arquivo da área de interesse desejado. Essa configuração é realizada por uma interface própria no modelo da OSCV em ambiente designado ao administrador, no qual encontra-se a opção de configuração dos locais de extração, como endereço *web* e as expressões utilizadas para navegação até a localização de fontes especificadas. No caso de haver alguma alteração no portal ou alguma ocorrência de exceção na tarefa de extração, ou ainda, alguma modificação no formato do arquivo que passará a disponibilizar a partir daquele momento, o sistema emitirá notificações para o administrador realizar ajustes. Essa atividade acontece para que não haja interrupções na etapa de atualização do *data warehouse*. As notificações serão enviadas por e-mail aos responsáveis cadastrados na OSCV.

Para garantir que o *data warehouse* contenha sempre dados atualizados, configura-se períodos de tempo em que a extração será executada. Essa configuração é efetuada por meio de agendamento prévio dentro da ferramenta de *crawling*. Com isso, a rotina de coleta será realizada automaticamente, sem que haja a intervenção do administrador, evitando possíveis falhas humanas. Ao realizar a configuração das

fontes desejadas inicialmente, já se conhece o formato que as mesmas irão disponibilizar seus arquivos, para, consequentemente, programar-se o script ETL de carga do *data warehouse*. Os scripts ETL realizam a leitura dos conjuntos de dados das fontes de interesse em seu estado bruto, executam atividades de saneamento e transformação, e por último, armazenam os dados extraídos nas tabelas.

4. Resultados Preliminares e Considerações Finais

A OSCV é uma arquitetura projetada para que o usuário possa usufruir dos dados públicos disponíveis pelos órgãos governamentais de maneira mais clara e facilitada. A contribuição deste trabalho é a automatização na coleta de dados, que garante a atualização das informações armazenadas no *data warehouse* da ferramenta, abstraindo tempo de manuseio de administradores e garantindo a coleta de informações recentes no portal de transparência em períodos de tempo pré-determinados. A automatização pode viabilizar o crescimento da base de dados da OSCV e seu potencial de abrangência enquanto ferramenta de acesso à informação pelos cidadãos, contribuindo ainda mais para o desenvolvimento de cidades inteligentes.

Para avaliar o processo desenvolvido de automatização, será realizado uma futura etapa, na qual será disponibilizado o acesso a ferramenta para que determinado número de usuários possa testá-la. Serão aplicados questionários de avaliação de usabilidade e aceitação para os usuários participantes. Os questionários serão produzidos com os métodos SUS (*System Usability Scale)* [Brook 1996] e TAM (*Technology acceptance model)* [Venkatesh *and* Bala 2008]. Analisado os resultados obtidos, será possível apontar melhorias e ajustes necessários, que posteriormente serão incorporados na ferramenta, atendendo assim os objetivos propostos no trabalho.

Referências

Avelar, R. E. A. 2013 Cidades inteligentes: uma abordagem tecnológica. v. 12.

Berst, J. 2013. Smart Cities Readiness Guide. Redmound, WA, USA.

Brook, J. (1996). SUS: a quick and dirty usability scale. *Usability Evaluation in Industry*. London: Taylor & Francis. p. 189–194.

Gray, J.;Dietrich, D.;Mcnamara, T. 2010. Manual dos dados abertos: governo. 58p.

Lusa, D. A. 2016. Open Smart City View - Uma Solução Computacional para Manipulação e Apresentação de Dados Governamentais Abertos. Biblioteca de Dissertações da Universidade de Passo Fundo.

Schuurman, D.;v Baccarne, B.; De Marez, L. 2012. Smart Ideas for Smart Cities. *J. of theoretical and applied electronic commerce research*, v. 7, n. 3, p. 11–12.

Venkatesh, V. e Bala, H. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, v. 39, n. 2, p. 273–315.