Mineração da *web* para estimação e monitoramento de índices de preços de imóveis de Florianópolis

Rafael Bassegio Caumo[1], Henrique Lopez Blanck[2], Alexandre Leopoldo Gonçalves[3], José Leomar Todesco[4], João Artur de Souza[5]

ABSTRACT

Digital databases - including those also known as big databases - are rich and yet little explored as knowledge sources. Every day, with the advancement of the digital age and the popularization of the internet, quantities and varieties of new data are generated, stored and made available with speed. As a central element, the Internet as a Data Source is an emerging concept that has been perceived as a competitive gain opportunity for companies. This concept has also been proved as efficient to the decision-making process for construction and implementation of public policies. In this context, this article aims to explore Knowledge Engineering techniques on the transformation of digital data available on the internet into knowledge for guiding decision-making processes of agents involved with the residential real estate market and its socioeconomic developments. The proposal collects – trough a web crawler – for sale residential real estate ads from Florianópolis, Brazil, and provides price indexes and related statistics in an interactive visualization tool – a system titled "Tá bombando, querido!".

Keywords: Web Mining; Internet as a Data Source, Web Crawler; Real Estate Price Index.

RESUMO

Bases de dados digitais, inclusas as também conhecidas como bases de *big data*, são ricas e ainda pouco exploradas como fontes de conhecimento. A cada dia, com o avanço da era digital e a popularização da internet, quantidades e variedades de novos dados são gerados, armazenados e disponibilizados com velocidade. Como um dos elementos centrais, a *Internet as a Data Source* é um conceito que surge e tem sido percebido como oportunidade de ganho competitivo para empresas e de suporte para o processo de tomadas de decisão no âmbito de políticas públicas. Nesse contexto, este artigo se propõe a explorar técnicas de Engenharia do Conhecimento para a transformação de dados digitais à disposição na internet em conhecimento que possa orientar processos de tomadas de decisão de agentes envolvidos com o mercado imobiliário residencial e seus desdobramentos socioeconômicos. A proposta coleta – via *web crawler* – anúncios de imóveis residenciais oferecidos para venda na cidade de Florianópolis, Brasil, e disponibiliza índices de preços e estatísticas relacionadas em uma ferramenta interativa de visualização – um sistema intitulado "Tá bombando, querido!".

Palavras-chave: Mineração da *Web*; *Internet as a Data Source, Web Crawler*; Índice de Preços Imobiliários.

[1] Doutorando em Engenharia e Gestão do Conhecimento na Universidade Federal de Santa Catarina, Brasil. E-mail: rbcaumo@gmail.com

[2] Mestrando em Engenharia e Gestão do Conhecimento na Universidade Federal de Santa Catarina, Brasil. E-mail: henriqueblanck@gmail.com

[3] Professor do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina, Brasil. E-mail: a.l.goncalves@ufsc.br

[4] Professor do Departamento de Engenharia do Conhecimento da Universidade Federal de Santa Catarina, Brasil. E-mail: jose.todesco@ufsc.br

[5] Professor do Departamento de Engenharia do Conhecimento da Universidade Federal de Santa Catarina, Brasil. E-mail: jartur@gmail.com

1 INTRODUÇÃO

Indivíduos e organizações que produzem conhecimento a partir de dados tem se deparado com uma recente e vasta disponibilidade de matéria-prima para suas análises: os dados digitais. Produto do fenômeno da revolução dos dados, os dados digitais são gerados a partir dos rastros da interação humana com dispositivos digitais – como a internet e os telefones móveis – ou da captura de eventos cotidianos realizada por sensores e *scanners* – impulsionados pela Internet das Coisas (Kitchin, 2014).

Como alguns exemplos, dados digitais são registros armazenados em aplicativos e ligações telefônicas, mensagens e posicionamento de telefones móveis, interações em redes sociais virtuais, páginas da internet, registros em mecanismos de buscas, sensores científicos, de tráfego, de segurança, medidores inteligentes, imagens de satélite, rastreamento por GPS, transações comerciais, financeiras e bancárias, registros administrativos de serviços públicos (hospitais, programas sociais, etc.), entre outros (Kitchin, 2014).

Bases de dados derivados de tais registros, também tratadas simplesmente por *big data*, são fontes baratas, abundantes e pouco exploradas de conhecimento. Para empresas, saber extrair conhecimento a partir desse tipo de dado pode representar diferencial competitivo. Representa oportunidade de baixo custo para conhecer melhor o mercado. Para estatais, é base para o planejamento de políticas públicas. É oportunidade para contornar problemáticas – substituindo ou complementando – dos métodos tradicionais de produção de estatísticas públicas/oficiais e indicadores socioeconômicos (Vaccari, 2014; Kitchin, 2015).

Assim, dados digitais podem servir de insumo quando da construção de um sistema de conhecimento – a partir de metodologias de engenharia do conhecimento (Schreiber et al., 2000) – permitindo que organizações possam ampliar seus ativos de conhecimento, agregando valor à instituição (Mayer-Schonberger & Cukier, 2013).

No contexto da exploração dos dados digitais – especificamente daqueles advindos das páginas de internet, aproveitando-se da *Internet as a data Source* (European Commission [EC], 2012; Askitas & Zimmermann, 2015) – para descoberta e criação de conhecimento, o artigo se propõe a realizar uma aplicação prática que interesse tanto à instituições ou agentes privados que de alguma maneira se relacionam com o mercado imobiliário residencial quanto à organizações produtoras de indicadores socioeconômicos e estatísticas públicas e oficiais sobre o setor, apresentando as descobertas metodológico-operacionais.

Mais especificamente, o objetivo deste trabalho é explorar técnicas de engenharia do conhecimento para a transformação de dados digitais à disposição na internet em conhecimento

que possa orientar processos de tomadas de decisão de agentes envolvidos com o mercado imobiliário residencial e seus desdobramentos socioeconômicos. Nesse contexto, perpassa desde a coleta até o desenvolvimento de um sistema de conhecimento em forma de ferramenta de visualização, voltado para o setor imobiliário residencial de Florianópolis, que permite a identificação e o acompanhamento das áreas de maior e menor valorização imobiliária no município. Para tanto, tratando dados digitais derivados do conteúdo de páginas da internet como matéria prima em um processo de aquisição automática de conhecimento (Turban, Aronson & Liang, 2004).

A operacionalização consiste na captura, com técnicas de mineração na *web*, de dados de anúncios relativos à imóveis residenciais de Florianópolis em páginas da internet, com tratamento destes dados, geração de estatísticas, indicadores, gráficos e mapas e disponibilização de todo o conteúdo em um sistema *web based* que possui estrutura interativa de visualização de dados.

Assim, acredita-se que a contribuição predominante do artigo está na integração de técnicas, métodos e ferramentas de engenharia do conhecimento em uma aplicação prática, com resultados e propósitos que aqui estão inseridos no contexto do mercado imobiliário residencial, mas que podem ser extrapolados para outros contextos e domínios de interesse.

Após esta introdução, o artigo segue com a fundamentação teórica (Seção 2), a apresentação dos procedimentos metodológicos utilizados (Seção 3), o sistema de conhecimento construído (Seção 4), e por fim, algumas considerações finais (Seção 5).

2 REFERENCIAL TEÓRICO

2.1 REVOLUÇÃO DOS DADOS

Atores que constroem ou descobrem conhecimento a partir de dados como forma de orientar seus processos de tomadas de decisão, assumindo um comportamento *data-driven*, deparam-se com um contexto de oportunidade de ampliar suas fontes de matéria-prima (Askitas & Zimmermann, 2015).

O contexto está inserido no ambiente da revolução digital, acompanhada do avanço tecnológico e da popularização da internet. Com a explosão na utilização de dispositivos digitais ligados à internet, percebe-se o surgimento de uma nova classe de dados: os digitais. Estes derivam dos rastros da interação humana com dispositivos digitais, como a internet e os

telefones móveis, ou da captura de eventos cotidianos por sensores e scanners (World Economic Forum [WEF], 2010).

Caracterizados por bases com alta variedade temática e estrutural, de grande volume, geradas com velocidade, os dados digitais, por vezes tratados simplesmente como *big data*, como faz UNECE (2013) ao propor uma taxonomia de classificação. São produtos de registros armazenados em aplicativos e registros de ligações, mensagens e posicionamento de telefones móveis, redes sociais virtuais, páginas da internet, mecanismos de buscas na *web*, sensores científicos, de tráfego, de segurança, medidores inteligentes, imagens de satélite, rastreamento por GPS, transações comerciais, financeiras e bancárias, registros administrativos de serviços públicos (hospitais, programas sociais, etc.), entre outros (Kitchin, 2014).

A quantidade de dados digitais, ou *big data*, sendo produzidos tem experimentado uma recente explosão, ocasionando o que está sendo chamado de "avalanche de dados" (Miller, 2010). Em dados mais recentes, percebe-se que o nível de crescimento do quantitativo de dados digitais no mundo está na ordem de 100% ao ano (Helbing et al., 2016), fazendo com que a disponibilidade por dados digitais nos dias atuais supere a de dados analógicos ou mecânicos, também conhecidos por "*small data*", conforme comparação feita por Kitchin (2015).

O processo de transformação da predominância de disponibilidade por dados analógicos e *small*, produzidos predominantemente por levantamentos amostrais, para a dos dados digitais e *big data* compõem o contexto do que está hoje sendo chamado de Revolução dos Dados (Kitchin, 2014).

Os dados digitais, produtos da revolução dos dados, já vêm tendo seu potencial percebido científica e economicamente, sendo tratados como o "novo petróleo" (WEF, 2010; Bossoi, 2014), uma vez que oferecem a perspectiva de registros sobre os mais diversos aspectos da vida dos indivíduos de uma população, viabilizando a investigação e a construção de conhecimento a respeito de desde aspectos sociais até análises econômicas de mercado. Mayer-Schönberger e Cukier (2013) já previam que este fenômeno estaria pronto para "chacoalhar" tudo à medida que impacta na forma como o conhecimento é produzido, os negócios são conduzidos e a governança é promulgada.

2.2 MINERAÇÃO DE CONTEÚDO NA *WEB*

No âmbito da exploração da *Internet as a Data Source*, tida como uma das mais interessantes fontes associadas à revolução dos dados (EC, 2012; Askitas & Zimmermann, 2015), subsidio de estudos e análises em diversas áreas do conhecimento conforme exemplos

trazidos por Beresewicz (2015), as técnicas de mineração da *web* se apresentam como elemento central.

A mineração da *web* é descrita por Cooley et al. (1997) como o processo de descoberta e análise de dados e informações úteis na *web*, podendo ser subdividida em três tipos: mineração de conteúdo; mineração de estrutura; e mineração de uso.

No caso da mineração de conteúdo, consiste no trabalho de informações disponíveis dentro de documentos e páginas da *web*. A mineração de estrutura, por sua vez, considera as informações que detalham os links e os relacionamentos entre os objetos presentes na *web*. Enquanto a mineração de uso, por sua vez, trabalha com a informações deixadas na forma de rastros digitais por parte dos usuários ao navegarem por links, páginas e documentos da *web* (Cooley, 1997).

Em geral, a mineração na *web* é conduzida por um software computacional do tipo rastreador da rede, também conhecido como *web crawler*, que navega pela *web* de uma forma metódica e automatizada (Cheong, 1996) e podem ser do tipo generalistas ou *focused* (Jinbo et al., 2014; Chakrabarti et al., 1999; Fangfang & Xinwei, 2010).

2.3 VISUALIZAÇÃO DE DADOS

As técnicas de visualização de dados permitem estruturar representações de informações as quais podem ilustrar, por exemplo, padrões complexos, uma vez que os objetos visuais se completam em dimensões e cores. Essas representações também podem fornecer recursos de manipulação avançada dos dados para cortar, girar, ampliar os objetos ou efetuar recortes temporais para fornecer diferentes níveis de detalhes (Eick & Wills, 2003)

Na visualização, os conceitos de dados, informações e conhecimento, são utilizados muitas vezes em um contexto inter-relacionado. Em muitos casos, eles indicam diferentes níveis de abstração, compreensão ou veracidade. A visualização de dados, no contexto de descoberta do conhecimento, pode ser definida, segundo Fayyad, Grinstein e Wierse (2002), como uma ferramenta para obter informações sobre um espaço de informação, onde seu objetivo é a exploração de dados e informações de forma visual.

Para explorar informações a visualização de dados pode ser usada isoladamente ou em associação com outras tarefas, como análise de dependência, identificação de classe, descrição de conceito e detecção de desvio. Keim (2016) fornece uma elaborada análise de técnicas de

visualização para mineração de grandes quantidades de dados e classifica as técnicas de visualização em orientada a pixel, geometria de projeção e baseada em gráfico.

2.4 ÍNDICES DE PREÇO DE IMÓVEIS E ESTATÍSTICAS PÚBLICAS

Conforme Nadalin e Furtado (2011), a habitação é um bem econômico indispensável, sendo bem de consumo e/ou de investimento, com heterogeneidade no que se refere a seus atributos e localização. Este aspecto, por si só, constrói justificativa para a necessidade de se monitorar e analisar a variação no preço dos imóveis para um satisfatório processo de tomada de decisão por parte da demanda. Segundo o IBGE (2010), outras justificativas se inserem no contexto das estatísticas públicas ao colocar que índices de preços de imóveis possuem utilidade enquanto: indicador macroeconômico de inflação; indicador para política monetária; medida de riqueza; indicador (combinado a outros) de exposição ao risco; como deflator para as contas nacionais; e como indicador de comparação internacional.

Cabe reforçar que estatísticas públicas, adaptação nacional mais utilizada para o termo *official statistics*, são bens públicos essenciais para o desenvolvimento econômico, demográfico, social e ambiental de uma nação (UNECE, 1992). Tratam-se de insumo fundamental para planejamento e formulação de estratégias no mundo contemporâneo (Jannuzzi & Cracioso, 2002), seja nos setores público – subsidiando políticas e ações – ou privado – com análises de mercado e de identificação de oportunidades.

Hoje, no Brasil, alguns índices que envolvem preços de imóveis produzidos na forma de estatística pública são:

- Índice Geral do Mercado Imobiliário (IGMI), do Instituto Brasileiro de Economia (Ibre), da Fundação Getúlio Vargas (FGV), que estima a rentabilidade financeira de mercado com base na evolução da valorização dos preços do negócio imobiliário a partir de um levantamento amostral, apresentando série histórica de periodicidade trimestral que inicia em 2000;
- Índice Geral do Mercado Imobiliário Residencial (IGMI-R ABECIP), da Associação Brasileira das Entidades de Crédito Imobiliário e Poupança em parceria com o Instituto Brasileiro de Economia (Ibre), da Fundação Getúlio Vargas (FGV), que acompanha a evolução de preços de imóveis residenciais com base nos laudos de imóveis financiados pelos bancos e possui periodicidade mensal, com série histórica que se inicia em janeiro de 2014;

- FipeZAP, de parceria entre a Fundação Instituto de Pesquisas Econômicas (FIPE) e o portal ZAP Imóveis, que estima o preço médio do metro quadrado dos imóveis com base nos anúncios de imóveis, residenciais e comerciais, para venda e locação, registrados no próprio portal. A divulgação possui periodicidade mensal, com início em junho de 2012.

Além desses, outros atores também geram estatísticas sobre o mercado imobiliário no Brasil, como a Fundação Instituto de Pesquisas Econômicas Administrativas e Contábeis de Minas Gerais (IPEAD), que desde 1986 realiza e divulga pesquisas de aluguéis e lançamentos imobiliários, e alguns conselhos de corretores e sindicatos de habitação que utilizam-se de registros administrativos para gerar alguns indicadores, como o Conselho Regional dos Corretores de Imóveis do Estado de São Paulo (Creci-SP) e o Sindicato da Habitação do Distrito Federal (Secovi-DF).

Ou seja, possíveis bases dados para a construção de índices de preços imobiliários são: bases cartoriais; bases de bancos e financeiras que trabalham com empréstimos; bases administrativas de prefeituras; bases de pesquisas de sindicatos, conselhos e outras instituições; dados de anúncios disponíveis em imobiliárias e na internet; entre outras.

Percebendo a variedade de fontes, cada uma com diferentes prós e contras, conforme apresentam Nadalin e Furtado (2011), e também considerando os desafios metodológicos relacionados ao processo de geração de estimativas de qualidade para os índices de interesse, o IBGE (2010) se preocupou em fazer uma importante discussão que orienta a utilização de diferentes métodos já desenvolvidos para a geração de estimativas para índices de preços de imóveis, sendo eles: Método de Vendas Repetidas; Método de Avaliação; Método de Estratificação; e Método de Regressão Hedônica.

3 METODOLOGIA

No âmbito do paradigma quantitativo, o presente estudo utiliza dos princípios da pesquisa científica empírica de método indutivo e nível descritivo para desenvolver generalizações que contribuam compreensão de um fenômeno (Creswell, 1994; Gil, 2008), aqui representado pela evolução do preço do estoque de imóveis residenciais disponíveis para venda nos bairros de Florianópolis, utilizada como proxy para a valorização imobiliária das regiões. Além disso, contempla elementos de pesquisa tecnológica ao projetar e desenvolver um artefato do tipo sistema de conhecimento, do tipo *web based* e focado em técnicas de visualização

interativa, que tem como objetivo representar e apoiar atividades intensivas em conhecimento por parte de atores que trabalham com o mercado imobiliário residencial em Florianópolis e seus desdobramentos socioeconômicos.

Para construção das evidências empíricas, dados são coletados em pesquisa do tipo levantamento (Gil, 2008). A população alvo para o processo inferencial corresponde a todos os imóveis residenciais à venda no município de Florianópolis. Entretanto, o universo de pesquisa contemplou apenas os imóveis residenciais anunciados em uma única página na internet – aquela que apresentou maior quantidade de casos de imóveis na condição de interesse. Neste, os anúncios são coletados via mineração de conteúdo da *web* com auxílio de um *web crawler* desenvolvido em linguagem R, a partir do pacote *rvest* de forma automática e contínua (com periodicidade semanal).

O levantamento poderia ser considerado do tipo Censo ao passo que todos os registros de imóveis encontrados são coletados. Entretanto, o fato de se utilizar somente uma página da internet como fonte, gerando um provável erro de cobertura, faz com que o levantamento corresponda a uma amostra não probabilística do tipo intencional (Mattar, 1996), com um tamanho médio de 55 mil registros por coleta nos 20 primeiros processos de coleta de dados, realizados entre 13 de agosto e 24 de dezembro de 2017.

Após a coleta, é realizado um trabalho de pré-processamento dos dados, com eliminação de ruídos e tratamento de valores faltantes e discrepantes, ou seja, os *outliers*.

Alcançada uma base de dados estruturada e de qualidade, conduz-se um processo de inferência que pressupõe amostra probabilística não viesada em termos de representatividade ou seletividade, iniciando com a geração de estimativas para o preço médio por metro quadrado dos imóveis residenciais à venda em cada um dos bairros de Florianópolis, utilizando-se do Método de Estratificação (IBGE, 2010). Para delimitação dos bairros, é utilizada a mesma estrutura considerada nos sistemas de georreferenciamento da Prefeitura de Florianópolis, que contempla 37 bairros.

Por fim, índices e estimativas de preços são disponibilizados em um sistema de conhecimento que assume forma visual e interativa – com auxílio de técnicas de visualização implementadas pela combinação de aplicações *web* geradas na plataforma *Shiny* do *R project*, com desenvolvimento adicional em *HTML*, *CSS* e *Java Script*, a partir do framework aberto *Bootstrap* 4.

4 RESULTADOS

4.1 COLETA

As coletas dos dados relativos aos anúncios de imóveis foram realizadas de forma automatizada via *web crawler* construído em linguagem R com auxílio do pacote *rvest*. Foram iniciadas no dia 13 de agosto de 2017, repetidas uma vez a cada sete dias até 24 de dezembro de 2017. A fonte utilizada foi o site de anúncios que mais possui ofertas de venda de imóveis residenciais para o município de Florianópolis.

Dessa forma, foram coletados dados sobre todos os anúncios de imóveis residenciais - apartamento, casa, casa de condomínio, flat, sobrado, cobertura, *kitnet* - situados no município de Florianópolis. As informações capturadas sobre cada anúncio foram: endereço, título do anúncio, metragem, quantidade de quartos, quantidade de suítes, quantidade de banheiros, vagas de garagem, preço e uma descrição adicional do imóvel.

Cada realização da coleta gera uma base de dados que é adicionada à uma base geral e única, que contempla todas as coletas já realizadas.

4.2 PRÉ-PROCESSAMENTO

A etapa de pré-processamento dos dados é de extrema importância para garantir resultados válidos. Uma vez que os dados são mapeados e capturados, os algoritmos de pré-processamento são utilizados para executar verificações de validade e limpeza dos dados. Essa etapa é necessária para detectar ruídos (distorções, valores faltantes, valores discrepantes), bem como, para conhecer a estrutura das variáveis e perceber possibilidades de soluções analíticas, prevendo limitações e restrições.

No presente trabalho, as etapas de pré-processamento percorreram:

- Verificação e correção da ocorrência de registros duplicados;
- Filtragem de imóveis que correspondem à população de interesse de pesquisa;
- Remoção de caracteres não numéricos, corrigindo variáveis como Área e Preço;
- Cálculo de variáveis de apoio para o processo inferencial, como a variável Preço/m²;

- Construção da variável bairro, extraindo a informação da variável “endereço” através de lógica *Fuzzy*;
- Análise (tratando ou excluindo) de *outliers* e valores errados.
- Análise da estrutura da base final para antecipar problemáticas analíticas, como o problema das pequenas áreas.

4.3 ESTIMAÇÃO DOS ÍNDICES DE PREÇOS

De posse da base de dados validada, foram geradas as estimativas para o preço médio do metro quadrado por bairros de Florianópolis com a utilização do Método da Estratificação, descrito por IBGE (2010), Eurostat (2010) e Diewert (2007).

O Método da Estratificação, utilizado também pelo *Australian Bureau of Statistics* (ABS), consiste em subdividir os imóveis em estratos homogêneos dentro de cada bairro. Com o passar do tempo, a evolução dos preços médios dos estratos são verificadas e utilizadas para a composição de um índice de preço geral para cada bairro. O objetivo deste método é o de reduzir o problema causado pela qualidade (Diewert, 2007), isto é, corrige uma possível distorção na variação do preço que pode ser ocasionada pela alteração do perfil ou padrão dos imóveis de certa localidade no tempo.

A fórmula de cálculo do estimador do índice de preço de um bairro *i* pelo Método da Estratificação, utilizado como indicador da variação do preço médio do metro quadrado dos imóveis residenciais de Florianópolis entre os períodos 0 e *t*, é dada por

$$P_i^{0t} = \sum_{l=1}^{L} w_{i,l} P_{i,l}^{0t}$$

onde L é a quantidade de estratos do bairro *i*, $w_{i,l}$ é o peso do estrato l do bairro *i* e $P_{i,l}^{0t}$ é o índice do preço médio do metro quadrado no período *t* com base no período 0. Ou seja,

$$P_{i,l}^{0t} = \frac{P_{i,l}^{t}}{P_{i,l}^{0}}$$

onde $P_{i,l}^{t}$ e $P_{i,l}^{0}$ são os preços médios do metro quadrado no estrato l, bairro *i* e tempos *t* e 0, respectivamente.

Os pesos $w_{i,l}$, que podem ser calculados de diferentes maneiras, dependendo do objetivo da análise, foram aqui determinados pela proporção do estoque de cada estrato dentro do estoque total de imóveis do bairro. Os estratos, por sua vez, foram construídos com base nos cinco níveis da variável auxiliar “padrão do imóvel”, construída a partir de métodos de

mineração de dados – mais especificamente, aprendizagem não supervisionada via tarefa de agrupamento – análise de agrupamentos por particionamento através do algoritmo *k-means*.

Na prática, o resultado desta etapa é uma segunda base de dados que possui informações sobre os pesos e preços médios do metro quadrado de todos os estratos de todos os bairros para todos os períodos investigados.

4.4 FERRAMENTA DE VISUALIZAÇÃO

A partir da base de dados gerada foi possível construir o sistema de conhecimento final. Trata-se de uma ferramenta interativa de visualização estruturada na *web* (*web based*) e em tempo real, intitulada "Tá bombando, querido". Tem por objetivo central abastecer agentes tomadores de decisão, de alguma forma envolvidos com o mercado imobiliário residencial, com a explicitação de conhecimento sobre a evolução e o *status quo* dos preços nos bairros de Florianópolis.

O "Tá bombando, querido!" foi desenvolvido a partir da combinação de aplicativos *web* construídos no pacote *Shiny*, em linguagem R, com *layout* de página da internet estruturado em *Bootstrap* 4 - *Html*, *CSS* e *Java Script*. Está composto por um mapa, um gráfico, um painel de controle, instruções de uso e um *link* para um conteúdo de especificações metodológicas mais detalhadas, conforme demonstram as Figuras 1 e 2.

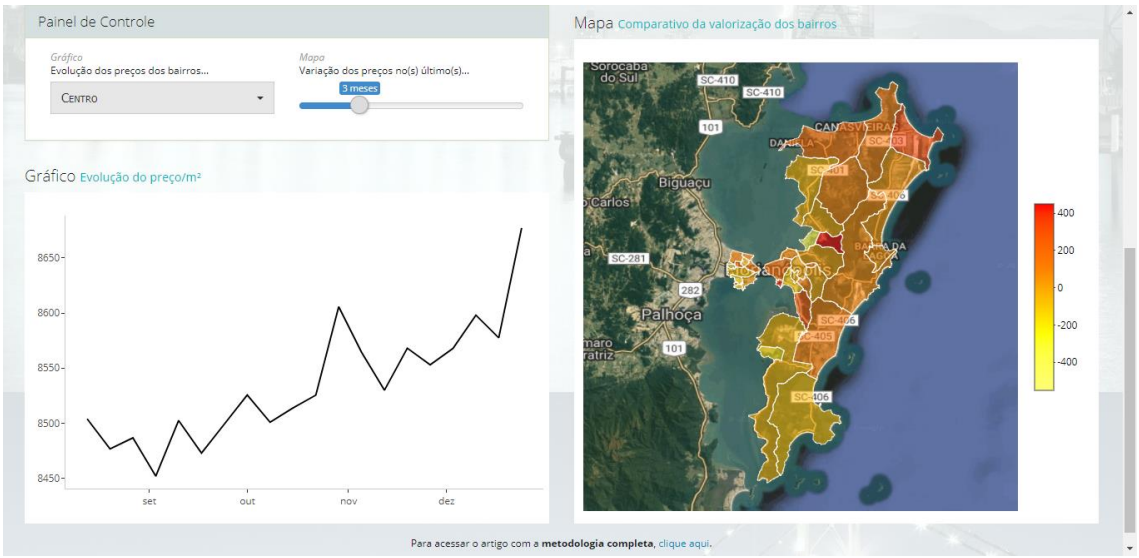
Figura 1 – Primeira parte da estrutura do sistema "Tá bombando, querido!"

Fonte: Elaborada pelos Autores

O mapa permite visualizar quais bairros tiveram maior valorização nos preços dos imóveis residenciais. Áreas mais escuras em vermelho tiveram maiores valorizações - ou menores desvalorizações - do que áreas em amarelo. O gráfico, por sua vez, permite observar como os preços médios do metro quadrado residencial estão evoluindo, possibilitando a comparação entre as séries históricas de diferentes bairros.

Ambos, mapa e gráfico, são interativamente controlados por controles situados no painel de controle. O primeiro dos controles permite selecionar até quatro bairros, apresentando no gráfico a comparação da evolução do preço médio por metro quadrado dos bairros selecionados. O segundo, por sua vez, permite que se defina qual o intervalo de tempo considerado na análise da variação dos índices de preços apresentados no mapa.

Figura 2 – Segunda parte da estrutura do sistema “Tá bombando, querido!”



Fonte: Elaborada pelos Autores

4.5 CONTRIBUIÇÕES

A exploração das técnicas de engenharia do conhecimento aqui realizada permitiu partir de dados digitais à disposição na internet para se chegar em um sistema interativo que disponibiliza informações que podem ser efetivadas em conhecimento de interesse para processos de tomadas de decisão de agentes envolvidos com o mercado imobiliário residencial e seus desdobramentos socioeconômicos.

Espera-se que, partindo da integração de técnicas, métodos e ferramentas aqui realizada, agentes de outros domínios de conhecimento possam viabilizar replicações, construindo suas

próprias aplicações práticas de interesse, ampliando as possibilidades de aproveitamento e de transformação em conhecimento dos dados digitais abertos disponíveis na internet.

5 CONSIDERAÇÕES FINAIS

Este artigo buscou alimentar o debate sobre a utilização de dados digitais, em especial de conteúdos de páginas da internet, como fontes para descoberta de conhecimento. O potencial e o contexto de oportunidades que se coloca com a revolução dos dados tem sito percebidos e imagina-se que cada vez mais os dados digitais vão substituir formas mais caras e operacionalmente complicadas de gerar e capturar dados e informações.

Para tanto, trabalhou-se com uma situação prática real no contexto do mercado imobiliário residencial. Como resultados, foram construídas soluções em termos de mineração de conteúdo *web*, via *web crawler*, e do tipo sistema de conhecimento, baseada na *web* e em visualização interativa.

O resultado parece trazer benefícios aos agentes tomadores de decisão envolvidos com o mercado imobiliário. Além disso, exemplificam e apresentam técnicas, métodos e ferramentas de engenharia do conhecimento que podem ser integradas na construção de aplicações práticas em outros domínios do conhecimento, ampliando as possibilidades de aproveitamento e de transformação em conhecimento dos dados digitais abertos disponíveis na internet.

Apesar de a transformação de dados digitais em conhecimento aqui realizada ter se demonstrada satisfatória, a literatura ainda reserva uma longa agenda de trabalhos no sentido de soluções para problemáticas metodológicas, tecnológicas, legislativas, financeiras, gerenciais e de privacidade associadas a geração de indicadores socioeconômicos a partir de dados digitais, conforme apontam Struijs et al. (2014), Tennekes & Offermans (2014) e Vaccari (2014).

No caso da aplicação em questão, a continuidade do índice de preços e, consequentemente, do sistema de visualização foi prejudicada em virtude da alteração na estrutura da página fonte dos anúncios imobiliários. Seis meses após o início da coleta, o controle da paginação foi modificado por parte do proprietário de maneira que inviabilizou que o *web crawler* percorra todos os anúncios. Em suas publicações, Tam & Clarke (2014), Daas (2015) e Kitchin (2015) discorrem justamente sobre a necessidade de cuidados com a certificação de que as fontes dos dados digitais não modifiquem a estrutura ou descontinuem o registro e a disponibilização das informações.

REFERÊNCIAS

Askitas, N., & Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, Vol. 36 Issue: 1, pp.2-12.

Bandyopadhyay, S., Maulik, U., Holder, L. B., Cook, D. J. (2005). Advanced Methods for Knowledge Discovery From Complex Data. *Advanced Information and Knowledge Processing.* London, U.K.: Springer-Verlag.

Beresewicz, M.. (2015). On the Representativeness of Internet Data Sources for the Real Estate Market in Poland. *Austrian Journal of Statistics.* 44. 45.

Bossoi, R. A. C. (2014). A proteção dos dados pessoais face às novas tecnologias. *Direito e novas tecnologias*, Florianópolis: CONPEDI.

Chakrabarti, S., Berg, M. V., & Dom, B. (1999). Focused Crawling: A New Approach to Topic-specific *Web* Resource Discovery. *Computer Networks.*

Cheong, F. C. (1996). Internet Agents. Spiders, Wanderers, Brokers and Bots (em inglês). Indianapolis: New Riders.

Creswell, J. W. (1994). Research design: Qualitative and quantitative approaches. Thousand Oaks, CA: *SAGE Publications.*

Daas, P. J. H., Puts, M. J., Buelens, B. and van den Hurk, P. A. M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*. 31(2), pp. 249-262.

Diewert, E. (2007). The Paris OECD-IMF Workshop on real estate price indexes: conclusions and future directions.

EC (European Commission). (2012). Internet as data source - Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. Recuperado em 15 de março, 2018 de: https://www.dialogic.nl/file/2016/12/2010.080-1226.pdf.

Eick, S. G., & Wills, G.J. (2003). Navigating large networks with hierarchies visualization.

EUROSTAT. Handbook on Residential Property Price Indices. Recuperado em 15 de março, 2018 de: http://epp.eurostat.ec.europa.eu/portal/page/portal/hicp/methodology/residential_property_p rice_indices.

Fangfang, X., & Xinwei, W. (2010). Analysis and comparison of *web* text clustering alogrithm, Computer Era, China.

Fayyad, U. M., Piatetsky-shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, CA, USA: *MIT Press*.

Fayyad, U., Grinstein, G.G., & Wierse, A. (2002). Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann.

Gil, A. C. (2008). Métodos e Técnicas de Pesquisa Social. São Paulo, *Atlas*.

Han, J., & Kamber, M.. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van den hoven, J., Zicari, R., & Zwitter, A. (2016). Behavioural Control or Digital Democracy? - *A Digital Manifesto.*

IBGE. (2010). Índice de preços imobiliários para o Brasil: estudos para discussão. 24p. Mimeografado. Recuperado em 15 de março, 2018 de: http://Www.Bcb.Gov.Br/Pec/Depep/Seminarios/2011_Iworkshopbcb/Arquivos/2011_Iworkshopbcb_Marlonsalazar.Pdf. 2010.

Jannuzzi, P. de M., & Cracioso, L. de S. (2002). Produção e disseminação da Informação Estatística pelas Agências Estaduais no Brasil. *Revista São Paulo em Perspectiva, Fundação SEADE*, V. 16 n.3.

Jinbo, W., Lianzhi, W., Wanlin, G., & Jian, Y. (2014). On an improved Naïve Bayesian keyword extraction algorithm, *Computer Application and Software*, China.

Keim, D.A. (2016). Visualization techniques for mining large databases: a comparison, *IEEE Transactions on Knowledge and Data Engineering*.

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructuresand Their Consequences. Sage, London.

Kitchin, R. (2015). Big Data and Official Statistics: Opportunities, Challenges and Risks. *Statistical Journal of the International Association of Official Statistics 31(3):* 471-481.

Mattar, F.N. (1996). Pesquisa de marketing. Sao Paulo: Atlas. 270p.

Mayer-schonberger, V., & Cukier, K. (2013). Big data: A Revolution that will transform how we live, work, and think. New York. *Houghton Mifflin Harcourt*.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science 50*, 181–201.

Nadalin, V., & Furtado, B.. Índices De Preços Para Imóveis: Uma Revisão.

Nickparsa, N. (2016). SpyBite: A New Approach to Designing a *Web* Crawler. California State University.

Schreiber, G. et al. (2000). Knowledge engineering and management: the CommonKADS methodology. Massachusetts: MIT Press. 471p.

Struijs, P., & Daas, P. J. H. (2013). Big Data, Big Impact? *Paper for the Seminar on Statistical Data Collection*, September 25–27, Geneva. Switzerland.

Tam, S-M. & Clarke, F. (2014). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. *Paper presented at International Conference on Big Data for Official Statistics*, Beijing, China, 28-30 Oct 2014.

Tennekes, M., & Offermans, M. (2014). Daytime population estimations based on mobile phone metadata. *Paper prepared for the Joint Statistical Meetings*, Boston.

Turban, E., Aronson, J., & Liang, T. (2004) Decision Support Systems and Intelligent Systems, Prentice Hall, New Jersey

UNECE. (1992). Fundamental Principles of Official Statistics in the UNECE region. Geneva.

UNECE. (2013). Classification of Types of Big Data. Recuperado em 15 de março, 2018 de: https://statswiki.unece.org/display/bigdata/Classification+ of+Big+Data. 2013.

Vaccari, C. (2014). Big Data and Official Statistics., *PhD Thesis*, School of Science and Technologies - University of Camerino.

WEF. (2010) Report about Personal Data: The Emergence of a New Asset Class, World Economic Forum.