EXTRAÇÃO DE DADOS DO SITE *TRIPADVISOR* COMO SUPORTE NA ELABORAÇÃO DE INDICADORES DO TURISMO DE MINAS GERAIS: UMA INICIATIVA EM *BIG DATA*

Emails:
rafalolbh@hotmail.com
renatabaracho@eci.ufmg.br

Rafael Almeida de Oliveira[1], Renata Maria Arantes Baracho Porto[2]

Resumo

A pesquisa tem como objetivo estudar o fenômeno *Big Data* e a possibilidade de utilização de ferramentas de extração de dados em ambiente *web* para auxiliar na elaboração de indicadores a respeito dos atrativos turísticos do estado de Minas Gerais cadastrados na maior plataforma de avaliação de destinos turísticos mundial denominada *TripAdvisor*. Para tanto, foi realizado um levantamento bibliográfico de autores ligados à ciência da informação e o papel das ferramentas de extração de informações via *web*. Após a contextualização, foi utilizada uma ferramenta de extração de dados denominada *Import.io* para coletar dados do site *TripAdvisor*, buscando as principais informações dos atrativos turísticos de Minas Gerais e transformando-as em um banco de dados estruturado. Assim, foi possível analisar informações como a oferta de atrativos por categorias no estado e por município, o número de avaliações, o perfil dos visitantes, o nível de satisfação e o período de maior visitação de cada um dos atrativos. Espera-se que a metodologia apresentada possa auxiliar o poder público estadual e municípios a elaborar indicadores de desempenho a partir da extração de dados já disponibilizados na *web*, com baixo custo, otimizando as ações e garantindo uma melhoria no uso de recursos públicos nas políticas relacionadas ao turismo.

Palavras-chave: Extração de dados. Turismo. B*ig data.* T*ripadvisor.* Recuperação da informação. estão da informação.

Abstract

The research aims to study the phenomenon called “Big Data” and the possibility of using free web data extraction tools (web scrapers) to help the development of indicators about tourist attractions in Minas Gerais State (Brazil) registered in the world’s most famous travel-related website known as “TripAdvisor”. Therefore, we carried out a brief study of themes such as information sciences and the role of web-based information extraction tools. After the literature review, we used a web scraper tool called Import.io to collect data from TripAdvisor, searching for key information of Minas Gerais’ tourist attractions and turning them into a structured database. Thus, it was possible to extract information such as the division of tourist attractions by categories from the state and municipalities, the number of evaluations, visitors' profiles,

[1] Mestrando em Ciência da Informação pela Escola de Ciência da Informação. Universidade Federal de Minas Gerais.
[2] Professora do Departamento de Teoria e Gestão da Informação. Escola de Ciência da Informação. Universidade Federal de Minas Gerais, Brasil.

satisfaction levels, and the period of most visits at each of the attractions. We expect this methodology to assist the state authorities and municipalities to create performance indicators form data extraction that is already available on the web at low cost, improving actions and ensuring an improvement in the use of public resources in tourism policies.

Keywords: Web scraping. Turism. Big data. Tripadvisor. Information retrieval. Information management.

1 INTRODUÇÃO

Presenciamos na sociedade moderna, um aumento do número de dados por cada pessoa ou organização no seu dia-a-dia. Com o avanço da tecnologia exemplificado pelas altas conexões de internet e utilização de dispositivos móveis, o acesso à informação se tornou mais fácil, gerando um volume de dados bem maior do que nas décadas anteriores. Cohen (2002), afirmou que vivíamos na era da "economia da informação" onde a forma de utilizar a informação por empresas e governos era decisiva para obter bons resultados econômicos.

Para alguns autores, uma das formas de auxiliar na interpretação e utilização dos dados está relacionada ao conceito de *Big Data*, ou seja, são conjuntos de dados gigantescos e muito complexos que necessitam de servidores, formas de gerenciamento, análises e tecnologias de visualização robustos e diferenciados (CHEN; CHIANG; STOREY, 2012), que não podem ser analisados por programas ou ferramentas de uso comum da informação (DAVENPORT, 2014).

O estudo do fenômeno informacional *denominado Big Data* vem ganhando a cada dia, maior importância no campo da Ciência da Informação. Ribeiro (2014; p. 102) cita que "a Ciência da Informação nasceu e foi gestada com o objetivo maior de apresentar solução para problemas ligados ao uso de dados e informação e como tal, tem um importante papel nos estudos que envolvem o tema *Big Data*" Para Boyd e Crawford (2012), estudos de *Big Data* são multidisciplinares, visto sua potencialidade de aplicação para diferentes áreas. Assim, pretende-se que esse trabalho possa ampliar o debate sobre o tema nas áreas correlatas tais como o Turismo.

Grande parte das informações características do fenômeno *Big Data* é disponibilizada de forma dispersa na internet, dificultando sua captura, tratamento e análise para tomada de decisões por gestores de organizações e empresas. Muitas dessas informações são disponibilizadas por diferentes usuários em redes sociais a partir de avaliações e comentários sobre produtos e serviços com o intuito de auxiliar demais usuários da web na solução de problemas.

Um dos exemplos pode ser caracterizado no setor de turismo pelo site *TripAdvisor.* A plataforma se caracteriza como a maior rede social para troca de informações sobre destinos turísticos no mundo e é alimentada pelos próprios usuários a partir da avaliação quantitativa e qualitativa de produtos e serviços turísticos, facilitando a escolha dos destinos e o planejamento de viagem dos turistas. De acordo com informações do próprio site (www.tripadvisor.com.br), as páginas de viagens do *TripAdvisor* atingiram em fevereiro de 2016, 340 milhões de visitantes únicos por mês, com 350 milhões de avaliações e opiniões cadastradas de mais de 6,5 milhões de meios de hospedagem, restaurantes e atrativos de 136 mil destinos.

Assim, acredita-se que o monitoramento e a avaliação do comportamento dos usuários nessa rede possam auxiliar na elaboração de indicadores, com possibilidade de contribuírem para

a criação de estratégias que visem à melhoria do atendimento e da satisfação dos visitantes em atrativos e destinos turísticos de Minas Gerais. Um dos métodos que auxiliam a organizar essa informação é o uso de ferramentas de extração de dados via web.

Sabe-se também que para a realização de pesquisas e levantamento de informações no setor público, são necessários recursos que na maioria das vezes, encontram-se escassos. De acordo com o Portal da Transparência do Governo de Minas Gerais , no ano de 2015, a Secretaria de Estado de Turismo de Minas Gerais (SETUR-MG) foi responsável por apenas 0,02% das despesas do Estado (R$ 16 milhões ante os R$ 80,79 bilhões), no qual mais da metade dessas despesas foram destinadas a custos relacionados à própria administração da Secretaria (tais como pagamento de vencimento dos servidores, diárias, materiais de consumo), sendo o recurso destinado à execução de projetos insuficiente para a elaboração de pesquisas de grande porte.

Esse presente trabalho pretende então, auxiliar na elaboração de indicadores de desempenho do turismo em Minas Gerais, a partir da extração de dados em redes sociais, mais especificamente o site *TripAdvisor*. Para tanto, é necessária a utilização de uma ferramenta de extração de dados que possibilite a análise de informações de forma clara.

Pretende-se também incentivar novas pesquisas relacionadas à gestão da informação no turismo, auxiliando os gestores públicos e privados na tomada de decisões, instigando-os a propor melhorias na coleta, monitoramento e divulgação das informações de forma estratégica. Isso possibilita benefícios não só para os órgãos governamentais que podem otimizar os recursos investidos, mas também para toda cadeia produtiva do turismo que terá informações mais precisas da atividade para a gestão de produtos e serviços turísticos.

2 BIG DATA

A utilização do termo "*Big Data*" foi observada aproximadamente no início da década de 1990, descrita pela *Association for Computing Machinery* simplesmente como um conjunto de dados muito grande que não poderia ser analisado por apenas um computador e que apenas nos últimos cinco anos, começou a ganhar popularidade (CRAWFORD; MITNER; GRAY, 2014), ampliando o seu conceito além do volume de informações.

McAfee e Brynjolfsson (2012) elencam três propriedades que são observadas apenas no conjunto de informações denominado como *Big Data*, denominadas como os "3 Vs", sendo o volume (grande fluxo de dados), velocidade (processamento em tempo real) e variedade (diversos formatos de dados). Alguns autores ainda acrescentam outra propriedade (ou um quarto "V") para *Big Data*. Han e Lu (2014), definem o "4º V" com a "veracidade" que reflete se os dados utilizados na análise comparativa estão de acordo com as características encontradas nos dados brutos. Já para a empresa Oracle (2015), o "4º V" seria dedicado ao "valor", ou seja, é necessária uma análise dos dados de forma coerente para que a informação seja repassada de acordo com as expectativas e de forma a auxiliar na resolução de problemas.

Grande parte das informações produzidas hoje é em formato digital, alimentada por diversos usuários e coletada e armazenada por computadores, sendo possível de ser organizada e extraída apenas por ferramentas computacionais (PUSCHMANN; BURGESS, 2014). Elas acabam dando significado aos bancos de dados que não conseguiriam ser compreendidos por um indivíduo ou um grupo de indivíduos. Portanto, criam-se ferramentas de "análise de *Big Data*". De acordo com Andrejevic "*Big Data* denota o momento onde formas automatizadas de

reconhecimento de padrões conhecidas como "análise de dados" se encontram com formas automatizadas de coleta e armazenamento de dados" (2014, p.3, tradução nossa).

As conceituações de *Big Data* defendida pelos autores apresentados permeiam não só os campos da informação e da computação, mas também abrem caminho para discussões sobre novos paradigmas na sociedade. Uma das definições que consegue consolidar esses pensamentos foi escrita por Boyd e Crawford (2012, p. 663) ao definirem todas essas características de *Big Data*:

> Nós definimos *Big Data* como um fenômeno cultural, tecnológico e acadêmico que repousa sobre a interação de: (1) Tecnologia: maximizando o poder computacional e a eficiência algoritma para reunir, analisar, relacionar e comparar bancos de dados volumosos. (2) Análise: lidar com grandes bancos de dados para identificar padrões com vistas a realizar reinvindicações sociais, técnicas e legais. (3) Mitologia: a crença generalizada que conjuntos de vários dados oferecem uma alta forma de inteligência e conhecimento que podem gerar "insights" que anteriormente eram impossíveis, com a aura da verdade, objetividade e assertividade (tradução nossa).

Essa definição apresenta-se como a mais indicada para a compressão do conceito e que norteará as demais discussões durante esse trabalho. Compreender a *Big Data*, não só como um problema computacional, mas entender que suas consequências podem de fato criar novos paradigmas para a sociedade é um ponto crucial que deve ser debatido com maior aprofundamento por diversas áreas do conhecimento.

No mundo dos negócios, o termo *Big Data* foi uma forma de se opor ao tradicional modo de levantamento de dados, sendo relacionado praticamente a uma "marca comercial". *Big Data* marcou uma mudança significativa entre os sistemas de gerenciamento de bancos de dados relacionais para plataformas que ofereciam vantagens de desempenho em longo prazo, ao contrário das soluções que eram encontradas no mercado (PUSCHMANN; BRUGESS, 2014).

Dentre as soluções tecnológicas disponíveis para auxiliar as organizações, a *Big Data* popularizou ferramentas de extração de dados, principalmente em ambiente *web*, com vistas a facilitar a organização das informações que elucidem a tomada de decisão.

3 EXTRAÇÃO DE DADOS WEB

Websites são criados na maioria das vezes para auxiliar na visualização de informações e não para exposição de dados de forma estruturada. Mas extrair essas informações de forma manual pode consumir muito tempo e ser suscetível a erros de extração. Nesse contexto, a extração das informações de forma automatizada possui um papel fundamental para a análise dos dados expostos. (DEVIKA; SURENDRAN, 2013).

Uma das possibilidades de extrair os dados dos websites é utilizar ferramentas denominadas *Web Scrapers* (ou apenas *Scraping*). De acordo com Vargiu e Urru (2013, p. 44, tradução nossa), um *web scraper* "foca em transformar dados não-estruturados da *web*, tipicamente em formato HTML, em dados estruturados que possam ser arquivados e analisados em bancos de dados locais [...]".

Para as mesmas autoras, o processo de *scraping* tornou-se essencial para a elaboração de pesquisas relacionadas ao comportamento social e cada vez mais são utilizadas no campo das ciências sociais. Complementando esse pensamento, elas afirmam que:

> [...] *scraping* faz nada menos do que destravar o "potencial sociológico" da internet: ele torna viável para as pesquisas sociais trabalhar com grandes quantidades de dados gerados pelo próprio usuário, que ultimamente estão se acumulando em plataformas online como por exemplo o Facebook, Twitter, Wikipedia, além de outros (MARRES; WELTEVREDE, 2012, p. 10, tradução nossa).

A extração de dados *web* possibilita então, analisar as informações em tempo real produzidas e alimentadas em plataformas digitais pelos próprios usuários e, compreender o comportamento desses usuários pode auxiliar diretamente a responder diversos questionamentos e problemas enfrentados diariamente por pesquisadores e organizações de diferentes setores, assim como o turismo.

4 TRIPADVISOR

Considerado o maior site de compartilhamento de informações de viagens do mundo, o *TripAdvisor* foi fundado no ano 2000 e desde então, se tornou uma corporação que administra e opera sites sob domínio de outras 24 empresas de viagens online, empregando em março de 2016, 3100 pessoas[3]. O site atua em 48 países e está disponível em 28 idiomas.

Dentre os recursos oferecidos pelo site, os usuários podem comparar preços e reservar hotéis, casas de temporada, voos e passeios de forma online, receber recomendações de destinos e meios de hospedagem a partir do perfil do usuário, publicar fotos de viagens, interagir através de fórum de perguntas e dúvidas para cada destino, buscar locais e atrações a partir de interação com mapas virtuais e principalmente avaliar os produtos e destinos turísticos a partir da distribuição de notas e comentários abertos.

Para Torres, Morales e Jiménez (2013), os comentários e opiniões dos usuários do *TripAdvisor* possuem não só componentes quantitativos (posições em rankings, avaliações, número de comentários) mas também contam com informações qualitativas, possibilitando a um proprietário de um estabelecimento, por exemplo, a identificar o que os seus clientes estão dizendo sobre seu estabelecimento e em caso negativo, poder corrigir falhas de forma rápida e precisa. O autor também afirma que:

> Processar as informações e opiniões dos clientes de forma correta possibilitará as empresas a melhorar os processos de produção, o que é uma grande valia para as empresas que sabem que ter todas as informações dos consumidores na ponta dos dedos e de forma gratuita é vantajoso (TORRES; MORALES; JIMÉNEZ, 2013, p. 23, tradução nossa).

Limberger *et al*. (2014, p. 62, tradução nossa) afirmam que é difícil categorizar o *TripAdvisor* como uma rede social, uma comunidade virtual ou *blog*. Porém, fica claro que o

[3] Fonte: *TripAdvisor* – www.tripadvisor.com.br . Acesso em: 9 de maio de 2016.

objetivo do site é "coletar e disseminar conteúdos elaborados pelos usuários sobre viagens [...], sendo que suas características mais marcantes são os comentários e avaliações".

Com o intuito de construção de conteúdo colaborativo, os próprios usuários do site cadastram os atrativos turísticos de cada destino, dividindo-os em categorias pré-determinadas, o que facilita na busca de informações através de filtros pelos interessados em conhecer um específico tipo de atrativo, como um museu.

5 METODOLOGIA

A pesquisa utiliza métodos quantitativos e qualitativos para realizar seu processo investigativo. O método quantitativo se baseia na utilização de instrumentos padronizados e neutros para a coleta de informações. Já o método qualitativo é mais indicado para investigações críticas e interpretativas, onde se deseja estudar fenômenos que envolvem diversas relações sociais entre indivíduos ou grupos de indivíduos, onde o investigador é considerado como instrumento primário e essencial na coleta e análise de dados (TEIXEIRA, 2003). Assim, considera-se que essa pesquisa possa ser caracterizada com elementos quantitativos e qualitativos, sendo denominada de híbrida (CRESWELL, 2003).

No caso desse estudo, a extração das informações, assim como o tratamento dos dados para a elaboração de indicadores serão tratados de forma quantitativa, buscando de forma censitária trabalhar com as informações dos atrativos e destinos de Minas Gerais cadastrados no site *TripAdvisor* que se referem à possibilidade de realizar cálculos numéricos, a partir da utilização de uma ferramenta de extração de dados *web* denominada *import.io*.

A escolha de utilização dessa aplicação para esse estudo foi realizada por ser uma ferramenta de extração de dados disponível de forma gratuita na *web*, além de possibilitar uma coleta ilimitada de dados e principalmente por não necessitar de conhecimentos de linguagem de programação para sua utilização, ou seja, foi considerada a ferramenta mais acessível para os propósitos da pesquisa.

O extrator possibilita ao usuário "ensinar" o aplicativo do *import.io* a extrair apenas aquelas informações que constam em uma determinada página (URL), sendo ela uma informação de texto, imagem, link ou número, registrando-as em um arquivo compatível com a utilização em banco de dados.

De forma geral, para coletar todas as informações selecionadas dos atrativos turísticos de Minas Gerais, foi realizado um trabalho de 3 etapas sendo a primeira para coleta de URLs com a listagem de atrativos de cada cidade, a segunda para coleta das URLs para cada um dos atrativos e a terceira para a extração das informações selecionadas.

Após a extração das informações, foi realizado o *download* do banco de dados para o *excel,* no qual foi necessário organizar as informações para posterior análise. No total, foram coletadas as informações de 1.324 URLs, sendo que 78 (5,9%) foram desconsideradas, pois se tratavam de URLs de grupos de atrativos ou serviços (e não de um único atrativo ou serviço), que são divulgadas nas páginas de alguns destinos e que redirecionam para outros atrativos. Além disso, caso as informações desses grupos de atrativos fossem extraídas, haveria uma grande repetição de informações de atrativos, portanto, para fins dessa pesquisa, optou-se por não extrair essas informações e retirá-las do banco de dados.

Dessa forma, foi considerado para a formação do banco de dados um total de 1.246 atrativos de 232 municípios cadastrados. Porém, percebeu-se que parte dos atrativos extraídos possuía um número muito pequeno de avaliações, o que dificultaria a análise de informações quantitativas como, por exemplo, o percentual de pessoas que avaliaram o atrativo como excelente, muito bom, regular, ruim ou péssimo. Sendo assim, decidiu-se por fazer um recorte no número de atrativos a partir do número total de avaliações de cada um (somatório das avaliações de excelente, muito bom, regular, ruim ou péssimo).

Para a realização desse recorte, tirou-se a média do número de avaliações dos atrativos. O resultado mostrou que, em média, cada atrativo possuía 96 avaliações e portanto, foram considerados para análise dos resultados os atrativos que possuíam número igual ou acima dessa média.

Após esse recorte foram considerados para a análise dos resultados 222 atrativos de 43 municípios, ou seja, 17% do número de atrativos e 18% dos municípios extraídos inicialmente.

6 RESULTADOS PRELIMINARES

Os resultados preliminares apontaram que foi possível coletar as informações de cada um dos atrativos selecionados e trabalhá-las de forma eficiente a partir de uma planilha do *excel* denominada "tabela dinâmica". Foi possível, dentre várias opções, identificar:

- a oferta de atrativos por município e por categoria (museus, igreja, shopping...);
- os atrativos, municípios e categorias com maior número de avalições absolutas, além de avaliações por segmento (excelente, muito bom...);
- o perfil dos visitantes para cada município, atrativo e categoria (família, sozinho, a negócios...)
- a época de visitação mais pertinente de cada município, atrativo e categoria.

Um dos exemplos de extração pode ser exemplificado pelo município de Mariana (Tabela 1). Pelos dados apresentados, percebeu-se que a categoria "minas" (*mines*) representa 34,58% do total de avaliações dos atrativos apresentados. Já a categoria "igrejas e catedrais" (*churches & cathedrals*) representa 32,15% de todas as avaliações. Foi possível observar também que de todos os atrativos do município, a maior nota de satisfação foi alcançada pela Praça Minas Gerais (4,5) e a menor pelo museu histórico Casa de Câmara e Candeia (3,9).

Tabela 1 – Exemplo de informações retiradas para o município de Mariana

Rótulos de Linha	avaliacoes	nota	familia%	casal%	sozinho%	negocios%	amigos%
Mariana	**2,01%**	**4,2**	**30,0**	**35,2**	**8,2**	**2,1**	**24,5**
Churches & Cathedrals	**32,15%**	**4,3**	**26,9**	**39,5**	**7,3**	**1,9**	**24,5**
Catedral Basilica da Se	35,37%	4,3	29,1	36,5	9,1	2,2	23,0
Centro Historico de Mariana	24,45%	4,2	23,4	40,0	4,1	1,4	31,0
Our Lady of Carmo church	16,74%	4,3	24,0	41,3	9,6	1,9	23,1
Sao Pedro dos Clerigos Cathedral	23,44%	4,2	31,0	40,0	6,2	2,1	20,7
History Museums	**6,93%**	**3,9**	**29,4**	**35,3**	**10,3**	**2,2**	**22,8**
Casa de Camara e Cadeia	100,00%	3,9	29,4	35,3	10,3	2,2	22,8
Mines	**34,58%**	**4,2**	**33,0**	**35,7**	**4,9**	**1,6**	**24,9**
Mina da Passagem	100,00%	4,2	33,0	35,7	4,9	1,6	24,9
Points of Interest & Landmarks	**15,49%**	**4,3**	**34,0**	**27,3**	**9,4**	**2,7**	**26,6**
Praca Gomes Freire	29,61%	4,2	34,4	25,6	12,2	4,4	23,3
Praca Minas Gerais	70,39%	4,5	33,6	29,0	6,5	0,9	29,9
Scenic Railroads	**10,86%**	**4,1**	**32,3**	**33,3**	**10,6**	**2,1**	**21,7**
Steam train to Ouro Preto	100,00%	4,1	32,3	33,3	10,6	2,1	21,7

Em relação ao perfil dos visitantes, observou-se que Mariana possui uma distribuição próxima nos perfis de pessoas que viajam a casal (35,2%), com a família (30,0%) e com amigos (24,5%).

Tabela 2 – Exemplo de informações retiradas para a categoria "museus especializados" em Minas Gerais

Rótulos de Linha	avaliacoes	nota	mar_mai%	jun_ago%	set_nov%	dez_fev%
Speciality Museums	**13,48%**	**4,3**	**18,1**	**27,5**	**26,7**	**27,7**
Inhotim	32,21%	4,8	19,9	26,1	26,0	28,1
Museum of Betrayal	11,09%	4,5	16,9	29,2	26,2	27,7
Memorial Minas Gerais Vale	10,41%	4,7	18,6	29,4	27,9	24,0
MM Gerdau - Museu das Minas e do Metal	6,78%	4,5	18,3	26,5	27,3	27,9
Palacio da Liberdade	5,87%	4,5	19,6	25,7	30,2	24,6
Museu De Artes & Oficios	3,59%	4,6	21,4	29,6	26,1	23,0
Museu de Sant'Ana	3,33%	4,5	14,0	24,9	29,7	31,4
Museum of Mineralogy	3,17%	4,5	16,5	29,9	29,5	24,2
Museu da Liturgia	2,76%	4,5	18,7	32,7	22,3	26,3
Oratory Museum (Museu Do Oratorio)	2,73%	4,5	18,4	28,1	26,1	27,4
Museu de Ciencia e Tecnica da Escola de Minas/UFOP	1,60%	4,5	16,6	30,1	26,6	26,6
Historical Museum of Dona Beja	1,56%	3,6	17,5	28,7	24,7	29,1
Museu Ferroviario	1,54%	4,3	18,6	24,9	28,1	28,5
Museu do Automovel	1,53%	4,1	19,5	32,3	21,4	26,8
Royal Road Automobile Museum	1,50%	4,2	12,1	32,1	27,9	27,9
Aleijadinho Museum	1,34%	4,2	22,9	25,5	26,0	25,5
Mariano Procopio Museum	1,34%	4,0	17,7	28,1	26,6	27,6
Museu De Arte Da Pampulha	1,33%	3,8	16,8	25,3	28,4	29,5
Estacao Ferroviaria de Ouro Preto	1,02%	4,1	22,6	24,0	30,8	22,6
Chico Xavier Memories and Reminders House	1,00%	4,6	24,5	20,3	21,7	33,6
Museu Historico e Geografico	0,95%	4,2	11,0	25,7	30,1	33,1
Casa dos Inconfidentes Museum	0,93%	4,5	23,3	21,8	21,8	33,1
Museu do Diamante	0,89%	3,7	18,1	29,1	25,2	27,6
Casa Guimaraes Rosa Museum	0,86%	4,4	17,9	22,0	27,6	32,5
Mariano Procopio Foundation Museum	0,69%	4,1	11,1	35,4	30,3	23,2

Outro exemplo é relacionado a análise de todos os museus de Minas Gerais que possuem o maior número de avaliações no site *TripAdvisor*, sendo que o Inhotim representa 32,21% de todas as avaliações realizadas (Tabela 2).

Em relação a satisfação dos usuários, percebeu-se que o Museu Dona Beja em Araxá alcançou a menor nota. Esse resultado auxilia diretamente o setor público e os gestores dos museus a identificarem onde se encontram os pontos de maior atenção para elaborar propostas de melhoria. Ao identificar problemas de satisfação no museu Dona Beja, foi realizada uma análise qualitativa dos comentários no site e percebeu-se que o museu ficou mais de um ano fechado para reforma, o que gerou insatisfação dos usuários. Em outro exemplo, no caso do Museu do Diamante em Diamantina, as maiores reclamações foram focadas na falta de infraestrutura e o baixo acervo. Assim, é possível criar políticas que auxiliem na melhoria das condições dos museus de forma pontual, visando a melhoria da satisfação dos usuários ao longo do tempo. Caso seja realizada uma reforma ou melhoria nas condições de visitação, o indicador poderá refletir numa melhoria das notas, possibilitando analisar de forma quantitativa o impacto de uma determinada ação de melhoria.

A extração possibilitou visualizar os períodos com o maior número de comentários dos viajantes, que se acredita ser em grande parte condizentes com as datas de visitação desses museus. Assim, é possível checar a possibilidade de criar roteiros integrados de visitação em museus que possuem uma taxa de avaliações baixas como por exemplo, em Dezembro a Fevereiro, criando iniciativas como programações gratuitas, maior divulgação dos espaços,

dentre outras ações, como é o caso do Museu de Mineralogia (24,2%), Memorial Minas Gerais Vale (24%), Palácio da Liberdade (24%) e Museu de Artes e Ofícios (23%), todos eles localizados em Belo Horizonte.

Tabela 3 – Exemplo de informações retiradas para os 10 municípios com maior percentual de perfil de visitantes com amigos ou casais

Rótulos de Linha	amigos%
Sao Thome das Letras	43,7
Lima Duarte	39,9
Conceicao da Ibitipoca	36,9
Serra do Cipo National Park	34,8
Brumadinho	33,7
Governador Valadares	33,5
Alto Caparao	33,3
Nova Lima	33,3
Santa Barbara	32,8
Capitolio	32,7

Rótulos de Linha	casal%
Goncalves	69,1
Lavras Novas	63,8
Monte Verde	57,6
Conceicao da Ibitipoca	43,4
Tiradentes	41,9
Carrancas	39,6
Serra do Cipo National Park	37,8
Sao Joao del Rei	36,9
Sao Thome das Letras	35,7
Monte Siao	35,3

Os resultados também mostram que os destinos com maior proporção de viagens de pessoas acompanhadas de amigos são característicos de natureza (cachoeiras, lagos, serras...) tais como São Thomé das letras, Lima Duarte (Serra do Ibitipoca), Conceição do Ibitipoca e Serra do Cipó (tabela 3). Já os destinos característicos de casais foram destacados em locais de pousadas, ligados a gastronomia e cultura, condizente com o que é divulgado para esse público como nos casos de Gonçalves, Lavras Novas e Monte Verde, cidades já conhecidas por serem atrativas para visitantes em lua de mel. Essas informações possibilitam elaborar propostas de roteiros ou divulgações desses destinos para um público específico, garantindo uma maior eficiência nas promoções realizadas pelo poder público.

7 CONCLUSÕES E PRÓXIMAS ETAPAS

As extrações realizadas possibilitaram a análise de diversas informações relevantes dos atrativos turísticos de Minas Gerais, tais como a oferta de atrativos por categorias e municípios, avaliação média de cada atrativo, perfil dos visitantes e época de maior visitação. Essas informações podem servir para gestores públicos e também de cada atrativo, monitorar o nível de satisfação dos visitantes, auxiliando na criação de projetos de melhoria da qualidade do serviço prestado, criação de roteiros ou divulgação de atrativos por segmentos e motivações de viagens. Desta forma, acredita-se que a extração de informações dos usuários no site *TripAdvis*or pode ser considerada como uma boa fonte de dados para planejamento do setor.

Na sequência da pesquisa, será necessária a realização de uma nova extração de dados com os mesmos atrativos, possibilitando comparar a evolução das informações ao longo do tempo, e posteriormente, propondo indicadores de desempenho do turismo.

Espera-se que a pesquisa possa futuramente, ser aprofundada, agregando novas metodologias de análise do banco de dados já existente ou utilizando novas extrações que possibilitem compreender mais sobre o perfil dos visitantes em Minas Gerais.

REFERÊNCIAS

ANDREJEVIC, Mark. The Big Data divide. **International Journal of Communication**, n. 8, 2014.

BARACHO, R. M. **A. Sistema de recuperação de informação visual em desenhos técnicos de engenharia e arquitetura:** modelo conceitual, esquema de classificação e protótipo. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

BOYD, Danah; CRAWFORD Kate. Critical questions for Big Data. **Information, Communication & Society,** ano 15, nº 5, p.662-679, 2012.

BRASIL. Ministério do Turismo. **Índice de competitividade do turismo nacional**: destinos indutores do desenvolvimento turístico regional: relatório Brasil 2014. Brasília, 2014.

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. **Business intelligence and analytics:** from Big Data to Big Impact, v. 36, n. 4, p. 1165-1188, 2012.

COHEN, Max F. Alguns aspectos do uso da informação na economia da informação. Brasília: **Ciência da Informação**, v. 31, n. 3, p. 26-36, 2002.

CRESWELL, J.W. **Research design:** qualitative, quantitative, and mixed methods approaches. 2. ed. Thousand Oaks: Sage Publications, 2003. 245 p.

DAVENPORT, Thomas. H. **Big Data no trabalho**. [S.l.]: Elsevier Brasil, 2014.

DEVIKA, K.; SURENDRAN, Subu. An overview of web data extraction techniques. **International Journal of Scientific Engineering and Technology**, v. 2, n. 4, 2013.

LIMBERGER, Pablo Flôres et al. Satisfaction in hospitality on TripAdvisor.com: an analysis of the correlation between evaluation criteria and overall satisfaction. **Tourism & Management Studies**, v. 10, n. 1, p. 59-65, 2014.

MARRES, Noortje; WELTEVREDE, Esther. Scraping the social? Issues in real-time social research. **Journal of Culture Economy** (subm), p. 1-52. Goldsmiths Research online, 2012. Disponível em: <http://eprints.gold.ac.uk/6768/>. Acesso em: 3 maio 2016.

ORACLE. **An Enterprise Architect's Guide to Big Data**. 2015. Disponível em: <http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf>. Acesso em: 19 abr. 2016.

ORGANIZAÇÃO MUNDIAL DO TURISMO. **Tourism Highlights**. [s.l.]: [s.n.]. 2007.

PUSCHMANN, Cornelius; BURGESS, Jean. Metaphors of Big Data. **International Journal of Communication**, nº 8, 2014.

RIBEIRO, Cláudio José Silva. Big Data: os novos desafios para o profissional da informação. **Revista Informação & Tecnologia**, v. 1, n. 1, p. 96-105, jan./jun. 2014.

TEIXEIRA, Enise Barth. A análise de dados na pesquisa científica: importância e desafios em estudos organizacionais. **Desenvolvimento em questão**, v. 1, n. 2, p. 177-201, 2003.

VARGIU, Eloisa; URRU, Mirko. Exploiting web scraping in a collaborative filtering-based approach to web advertising. **Artificial Intelligence Research**, v. 2, n. 1, 2012.