UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

BIANCA MOARA PIVATO MATT
VINÍCIUS DA SILVA MORAES

ANÁLISE DOS PERFIS DE TURISTAS DO TRIPADVISOR QUE VISITARAM CURITIBA

CURITIBA

2018

BIANCA MOARA PIVATO MATT
VINÍCIUS DA SILVA MORAES

ANÁLISE DOS PERFIS DE TURISTAS DO TRIPADVISOR QUE VISITARAM CURITIBA

Trabalho de Conclusão de Curso de graduação, apresentado à disciplina de Trabalho de Diplomação, do Curso Superior de Bacharelado em Sistemas da Informação do Departamento Acadêmico de Informática – DAINF – da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para obtenção do título de Bacharel em Sistemas da Informação.

Orientador: Thiago Henrique Silva

CURITIBA

2018

TERMO DE APROVAÇÃO

“ANÁLISE DOS PERFIS DE TURISTAS DO TRIPADVISOR QUE VISITARAM CURITIBA”

por

“Bianca Moara Pivato Matt e Vinicius da Silva Moraes”

Este Trabalho de Conclusão de Curso foi apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação na Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba. Os alunos foram arguidos pelos membros da Banca de Avaliação abaixo assinados. Após deliberação a Banca de Avaliação considerou o trabalho _____.

_____ **Prof. Thiago Henrique Silva** (Presidente - UTFPR/Curitiba)	_____ **Profa. Rita Cristina Galagarra Berardi** (Avaliadora 1 - UTFPR/Curitiba)
_____ **Prof. Alexandre Reis Graeml** (Avaliador 2 - UTFPR/Curitiba)	_____ **Profa. Leyza Baldo Dorini** (Professora Responsável pelo TCC – UTFPR/Curitiba)
_____ **Prof. Leonelo Dell Anhol Almeida** (Coordenador(a) do curso de Bacharelado em Sistemas de Informação – UTFPR/Curitiba)	

“A Folha de Aprovação assinada encontra-se na Coordenação do Curso.”

RESUMO

MATT, Bianca, MORAES, Vinicius. ANÁLISE DOS PERFIS DE TURISTAS DO TRIPADVISOR QUE VISITARAM CURITIBA. 56 f. – Curso de Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

Este trabalho teve por objetivo explorar as características dos usuários do TripAdvisor de acordo com as informações presentes em seus perfis nesta plataforma. Para tal, foram coletados e analisados os dados (mais especificamente cidades visitadas, *tags* informadas e selos concedidos) dos perfis de usuários do TripAdvisor que visitaram Curitiba e identificadas as relações entre os tipos de informações presentes nos perfis coletados. Além disso, buscou-se demonstrar a aplicabilidade dos resultados para possíveis utilizações em novos serviços. Como ferramenta de coleta utilizou-se dois *web crawlers* e os dados obtidos foram armazenados em um banco de dados não-relacional, o MongoDB. Os dados obtidos foram analisados com auxílio do algoritmo *Apriori*. Foram apresentadas possíveis aplicações dos resultados obtidos em diferentes etapas do processo de viagem, como a sugestão de destinos, sugestão de rotas, e planejamento de viagem, por exemplo.

Palavras-chave: TripAdvisor, Mineração de dados, Web Crawler

ABSTRACT

MATT, Bianca, MORAES, Vinicius. ANALYSIS OF TRIPADVISOR TOURISTS PROFILES WHO VISITED CURITIBA. 56 f. – Curso de Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

This thesis aimed to explore the characteristics of TripAdvisor users according to the information present in their profiles in this platform. In order to do so, we collected and analyzed data (more specifically visited cities, informed tags and given badges) from the TripAdvisor user profiles that visited Curitiba and identified the relationships among the types of information present in the collected profiles. In addition, it was demonstrated the applicability of the results for possible uses in new services. As a data collection tool, two web crawlers were used and the data obtained was stored in a non-relational database, MongoDB. The data obtained were analyzed with the aid of the textit Apriori algorithm. Possible applications of results obtained in different stages of the travel process, such as destination suggestion, route suggestion, and travel planning, were presented.

Keywords: TripAdvisor, Data Mining, Web Crawler

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O turismo, especialmente nos últimos anos, tem experimentado um crescimento mundial significativo (BATET et al., 2012). Segundo estatísticas de 2015 da United Nations World Tourism Organization (UNWTO), foi previsto um crescimento entre 3% e 4% no turismo internacional (UWNTO, 2014). Isso se deve, entre outras coisas, à enorme quantidade de informações que é possível encontrar na Web (BATET et al., 2012). As redes sociais online de turismo, a exemplo do *yelp.com*, *foursquare.com* e *TripAdvisor.com* (considerada a maior comunidade de viagens na Web (TRIPADVISOR, 2015)), têm transformado a forma como as pessoas organizam suas viagens. Cada vez mais os turistas, em busca de tomar melhores decisões durante o planejamento, têm se beneficiado e contribuído com revisões de hotéis, restaurantes e atrações nestes sites (AYEH et al., 2013). No momento de organizar uma viagem, essas informações aumentam o espectro de possibilidades de destinos e atrações para visitar, ao mesmo tempo que dificultam as escolhas devido ao grande número de opções que são apresentadas aos viajantes. A Web favorece também que serviços online sejam oferecidos a um custo relativamente baixo por parte dos negócios de turismo (KIM, 2004), o que colabora ainda mais para o crescimento da área.

O processo de planejamento de uma viagem pode ser dividido em estágios, como: escolher destinos, selecionar atrações turísticas, escolher acomodações, decidir rotas, entre outros (HUANG; BIAN, 2009). Neste cenário, é notável o papel e a importância de se aplicar recomendações ao turismo, em quaisquer etapas supracitadas, a fim de facilitar a tomada de decisões em relação a destinos, pontos turísticos e rotas, por exemplo, levando em conta as preferências dos turistas (BATET et al., 2012). Entretanto, identificar essas preferências depende, primeiramente, de se conhecer o perfil dos viajantes. Mesmo que diferentes autores em várias partes do mundo dediquem-se a estudos nesse sentido, compreender perfis se mostra uma tarefa bastante complexa, pois pode ser feita a partir de diferentes características, como será mostrado nas próximas seções deste trabalho. Por mais que, na maioria dos casos, a análise exploratória de perfis de usuários não seja o

objetivo final, é um passo essencial para que outros estudos sejam desenvolvidos, como é o caso dos métodos de recomendação. Sendo assim, neste trabalho foi realizada uma análise do perfil dos usuários do TripAdvisor em uma base de dados de colaboradores que avaliaram a cidade de Curitiba. Essa análise foi feita a partir das informações de Curitiba e das demais cidades avaliadas por eles e de atributos particulares desta comunidade online, especificamente *badges* (selos) e *tags* (etiquetas) de estilo de viagem, descritos detalhadamente na seção 3.1 sobre o TripAdvisor.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

Explorar as características dos usuários do TripAdvisor de acordo com as informações presentes em seus perfis nesta plataforma.

1.1.2 OBJETIVOS ESPECÍFICOS

- Coletar dados dos perfis de usuários do TripAdvisor que visitaram Curitiba.
- Analisar os dados provenientes dos perfis de usuários do TripAdvisor que visitaram Curitiba.
- Identificar relações entre os tipos de informações presentes nos perfis coletados.
- Demonstrar a aplicabilidade dos resultados para possíveis utilizações em novos serviços.

1.2 JUSTIFICATIVA

O desenvolvimento tecnológico, principalmente a partir da disseminação da Internet, dos sistemas wireless e da comunicação mobile, influenciou um novo tipo de comportamento entre empresas e consumidores. A partir do uso dessas tecnologias, as pessoas passaram a fazer buscas e comprar online, além de explorar e utilizar diversos aplicativos. O estilo de vida mudou e os benefícios vinculados à essa mudança, tais como eficiência, conveniência, redução de custos e diversidade, por exemplo, são evidentes (AMARAL et al., 2014).

Na área de turismo não foi diferente. Desde o final dos anos 90, a Internet tornou-se a principal fonte de informação utilizada pelos turistas. A partir de 2006, as redes

sociais, por permitirem a reunião de informações, postagem de comentários e compartilhamento de opiniões, passaram a exercer papel de grande influência no comportamento de compra dos viajantes (AMARAL et al., 2014). De acordo com Lu e Stepchenkova (2014), as plataformas de mídia social contêm grandes volumes de conteúdos gerados por usuários (*user-genterated content*) que são amplamente utilizados por turistas no momento de tomar decisões a respeito de uma viagem. Além disso, têm-se notado uma preferência entre os consumidores por obter informações por meio de ferramentas de busca online e nas redes sociais, em vez de meios de pesquisa tradicionais (LU; STEPCHENKOVA, 2014) (MUNAR et al., 2013). Exemplo disso é o que mostra a pesquisa de Gretzel et al. (2007), em que 96,4% dos 1.480 participantes revelaram utilizar a Internet como principal fonte de informações no momento de planejar uma viagem. Outras fontes tradicionais também foram citadas, mas com menor relevância, por exemplo, guias de viagem (68,3%), opiniões de amigos e familiares (41,6%), revistas (35,6%), panfletos (33,9%), jornais (27.8%), agentes de turismo (22,3%).

Ao planejar uma viagem, o turista busca alinhar aquilo que o destino oferece, tanto em termos de atrações quanto de acomodações, aos seus interesses e estilo de viagem. Reflexo disso é a crescente demanda do mercado por roteiros personalizados, que hoje chega a representar 70% das vendas de uma das maiores agências de viagem do Brasil (RUSSO, 2016). Segundo o presidente da Associação Brasileira de Agências de Viagem, Edmar Bull, em entrevista ao jornal Folha de S. Paulo, "o turista não quer mais um pacote embalado. A customização de viagens é uma tendência mundial e desde o ano passado [2015] vem crescendo no Brasil". Não somente isso, Bull afirmou ainda que "o brasileiro está mais exigente na hora de viajar"e que "ele quer dicas boas e quer fazer coisas diferentes"(RUSSO, 2016). A disponibilização de informações mais específicas e direcionadas facilita e acelera o processo decisório dos turistas, mas para tal é necessário conhecer o perfil de quem viaja. Isso vale, não só para o contexto das agências de viagem, mas essa realidade pode ser ampliada também ao mundo virtual.

Analisar o perfil dos usuários pode ser tarefa bastante complicada uma vez que, como já dito, são diversas as opções de aspectos a serem considerados. Em sites que permitem que sejam feitos *check-ins* é possível delinear o perfil de um usuário com base nos lugares frequentados por ele. Já em outros que disponibilizam alguma informação demográfica, essa classificação pode ser feita com base em idade e/ou gênero, por exemplo. Isso evidencia que a definição das características que serão analisadas pode variar em função do tipo de informação fornecida pela ferramenta com que se escolhe trabalhar. No caso deste estudo, optou-se pelo TripAdvisor, entre outros motivos, pelas informações

específicas disponibilizadas por esta plataforma, como os selos (*badges*) e as etiquetas (*tags*) de estilo de viagem, que serão explicadas detalhadamente no capítulo 3. Partiu-se da hipótese de que a análise destes atributos pode contribuir para um melhor entendimento do perfil dos usuários desta ferramenta e trazer *insights* diferenciados a respeito de suas preferências. As possíveis aplicações deste conhecimento se estendem da área de tecnologia, a partir do desenvolvimento de sistemas de recomendação ou de sugestão de rotas turísticas, até a realização de ações de marketing direcionadas, por exemplo.

Tendo isso em vista, é possível perceber como as redes sociais facilitam o consumo de informações por parte de seus usuários. Porém este meio virtual só existe uma vez que, além de consumidores, os próprios usuários são também os fornecedores dos conteúdos que circulam na rede. Por esta razão, além de melhor compreender o perfil dos usuários do TripAdvisor, outro fator importante no contexto deste trabalho é entender o que leva um colaborador a contribuir com informações, ao invés de apenas consumi-las.

Segundo Wasko e Faraj (2005) e Burkle et al. (2009), tal motivação ainda não é bem compreendida. Em seus trabalhos, que serão melhor apresentados no capítulo 2, os autores buscam explicar e fundamentar razões que induzem usuários a contribuir com informações em plataformas sociais. Wasko e Faraj (2005) citam exemplos como: ver outros colegas contribuindo, receber *feedbacks* positivos, ter uma grande audiência e o tópico em questão ser de interesse. Já Burkle et al. (2009) dizem que ter uma experiência a compartilhar e estar estruturalmente inserido na rede são fatores que podem contribuir para que os usuários se sintam motivados a fornecer informações. Porém, um tópico comum citado em ambos os trabalhos diz respeito à reputação. Segundo os autores, as pessoas compartilham seu conhecimento, além dos motivos supracitados, para aumentarem sua reputação na rede em que participam. Sabendo disso, os responsáveis pelas plataformas online criam mecanismos de incentivo a essa prática, afinal, suas ferramentas dependem do conteúdo gerado pelos próprios usuários. Olhando para o caso específico do TripAdvisor, podemos reconhecer o sistema de recompensas como um exemplo de mecanismo que alimenta a busca por reconhecimento e boa reputação. Esse sistema e seu funcionamento serão descritos em detalhes no capítulo 3.

Dito isso, o restante deste trabalho está organizado da seguinte forma: o capítulo 2 apresenta alguns dos trabalhos relacionados ao tema. O capítulo 3, Contextualização, apresenta os conceitos sobre as tecnologias utilizadas neste trabalho, como uma visão geral sobre o TripAdvisor, *web crawlers*, bancos de dados não relacionais, e mineração de dados. Já capítulo 4, Metodologia, aborda os métodos utilizados para o desenvolvimento

deste trabalho, bem como a forma como foram aplicados. Por fim, os capítulos 5 e 6 apresentam, respectivamente, os resultados obtidos e aquilo que foi analisado neste trabalho, assim como uma breve conclusão a respeito daquilo que foi discutido até agora e os possíveis caminhos para trabalhos futuros.

2 REVISÃO DE LITERATURA

São inúmeras as pesquisas realizadas na área que intercede o turismo e a tecnologia. Sistemas computacionais podem ser aplicados em diversos contextos e para diversos fins relacionados ao turismo. Neste capítulo são trazidos exemplos dessas aplicações, que serviram de base e inspiração para este estudo. Além disso, de forma complementar, julgou-se importante abordar trabalhos que discutem as motivações para usuários contribuírem com conteúdos em plataformas sociais online, bem como estudos que abordam o comportamento dos usuários nesses sistemas. Sendo assim, o capítulo foi dividido em quatro seções que abordam, respectivamente, a geração de possíveis rotas turísticas de acordo com pontos de interesse em uma cidade; a criação de sistemas de recomendação a partir da identificação dos interesses e classificação dos usuários; a análise do comportamento de usuários; e as motivações deles para contribuição nas redes sociais.

2.1 GERAÇÃO DE ROTAS TURÍSTICAS

No que tange a criação de rotas turísticas, o artigo de Brilhante et al. (2013) propõe uma nova estrutura para o planejamento turístico personalizado, utilizando, para isso, dados do Flickr. A ferramenta desenvolvida, chamada de TripBuilder, obtém informações sobre os itinerários seguidos por turistas diversos, e as combina com itinerários referentes ao ponto de interesse turístico disponível no Wikipédia. A tarefa de planejar rotas turísticas personalizadas é modelada como uma instância do problema de cobertura máxima generalizada. A inteligência coletiva dos usuários foi utilizada para derivar planos turísticos que maximizam uma medida de interesse para o turista, dadas suas preferências e orçamento. Assim como Brilhante et al. (2013), outros exemplos de trabalhos que utilizam dados do Flickr são os de Choudhury et al. (2010) e Majid et al. (2012). Neles as fotos da rede social são utilizadas para gerar itinerários turísticos de forma automática.

Há também estudos que utilizam dados provenientes de plataformas diferentes. Yoon et al. (2010) propõem uma arquitetura para recomendar rotas a turistas levando em

conta seus interesses e tempo de estadia. Yerva et al. (2013) apresentam um sistema de recomendação de itinerários baseado em preferências dos usuários, a partir de dados do Lonely Planet, Foursquare, e Facebook. Já Roy et al. (2011) criaram uma aplicação em que os próprios usuários constroem seus itinerários interativamente. Isso é feito a partir de *feedbacks* que consideram interesses pessoais e tempo disponível.

2.2 SISTEMAS DE RECOMENDAÇÃO

Sistemas de recomendação, por sua vez, podem ser aplicados em diferentes etapas de um plano de viagem, dentre eles, a escolha dos destinos, seleção de atrações turísticas e escolha de acomodações. O estudo de Bidart (2015), realizado com base no TripAdvisor, tem foco no estágio de escolha de destinos, e consiste em recomendar um conjunto de cidades para o usuário. Para tal, foram desenvolvidas duas estratégias para recomendar cidades utilizando filtragem colaborativa (BIDART, 2015). A primeira, baseada em vizinhança, explora a co-ocorrência de cidades visitadas pelos usuários e as avaliações das atrações das cidades feitas por eles. A segunda, baseada em fatores latentes, explora o texto das revisões feitas pelos usuários acerca das atrações das cidades e a posição geográfica das cidades como fator decisor para a recomendação.

Já os estudos de Cheng et al. (2011) e de Popescu e Grefenstette (2011) atuam na fase de escolha de atrações turísticas e propõem que as recomendações feitas aos usuários sejam personalizadas. O primeiro utiliza dados de diversos sites de compartilhamento de fotos, incluindo Flickr e Picasa, e sugere que as recomendações sejam personalizadas considerando perfis e atributos dos usuários, tais como idade, gênero e raça, por exemplo.

No segundo, por sua vez, que também utiliza dados do Flickr, essa personalização é feita com base na similaridade entre os destinos visitados pelos usuários. De acordo com o estudo desses autores, o processo é visto como um problema de filtragem colaborativa: são minerados os registros de lugares visitados pelos usuários e constrói-se uma matriz de similaridade usuário-usuário. Quando um usuário deseja visitar um novo destino, uma lista de atrações de visitantes potencialmente interessantes é produzida com base na experiência de usuários que já visitaram tal destino. Um trabalho que também utiliza similaridade, mas, dessa vez, entre usuários, é o de Zheng (2014), onde é apresentado um sistema de recomendação a partir de filtragem colaborativa e dados do TripAdvisor.

Outros exemplos são os trabalhos de Shi et al. (2011), que têm foco em recomendação de pontos turísticos utilizando dados do Wikipédia; de Baraglia et al. (2013),

que traz um modelo que prevê o próximo ponto de interesse de um turista com base em seu histórico; já o artigo de Pianese et al. (2013) teve como objeto de estudo *check-ins* do Foursquare a fim de descobrir pontos de interesse dos usuários com base em suas visitas.

2.3 COMPORTAMENTO DOS USUÁRIOS

Outra área de estudo relacionada a este trabalho aborda artigos que analisam o comportamento de usuários em redes sociais. Esta seção apresenta alguns deles.

O estudo de Cheng et al. (2011) fez uso de uma base de 22 milhões de *check-ins* no Twitter para encontrar um padrão de mobilidade. Os resultados mostraram que o comportamento dos usuários nesse sentido é influenciado por condições sociais, geográficas e econômicas. Autores como Hecht et al. (2011) também analisaram dados do Twitter para definir o comportamento dos usuários, mais especificamente o campo de localidade presente em seus perfis. O trabalho de Farrahi e Gatica-Perez (2011) teve como objetivo descobrir e analisar a rotina de pessoas, que segundo os autores, caracteriza comportamentos individuais e de grupos em termos de padrões de localização. Foram utilizadas informações provenientes de 97 telefones celulares em um período de 16 meses.

Já Lindqvist et al. (2011) utilizaram a base do Foursquare para analisar como e porque as pessoas usam serviços de compartilhamento de localização, além de discutir os problemas de privacidade referentes ao uso desse tipo de serviço. Ferreira et al. (2015) utilizam dados do Foursquare para analisar o comportamento de turistas e residentes em quatro cidades ao redor do mundo: Londres, Nova Iorque, Rio de Janeiro e Tóquio. Os autores descobriram que algumas localidades tem características mais relacionadas com o comportamento dos turistas e que, mesmo em lugares frequentados pelas duas classes, há uma distinção clara nos padrões de comportamento desses grupos. Long et al. (2013) investigam os *check-ins* de usuários do Foursquare na região de Pittsburgh, na Pensilvânia. Os resultados mostram que os viajantes exibem comportamentos diversificados em suas atividades. De modo similar, Pelechrinis e Krishnamurthy (2012) examinam as dinâmicas temporais dos usuários do que chamam de redes sociais baseadas em localização (do inglês *"location-based social network"* - LBSN). Os autores descobriram que usuários de diferentes LBSNs apresentam comportamentos diferentes em um lugar visitado dependendo do objetivo da rede que estão usando.

A seção a seguir traz estudos a respeito do que motiva usuários a contribuírem em redes sociais.

2.4 MOTIVAÇÃO PARA CONTRIBUIR

Os estudos de Wasko e Faraj (2005) e Burkle et al. (2009) abordam razões que ajudam a compreender e justificar o que motiva as pessoas a compartilharem informações por meio das ferramentas tecnológicas. Burkle et al. (2009) trazem um estudo voltado a novos participantes em uma rede social, no caso, o Facebook. Os autores partem de quatro hipóteses e se valem de três teorias sobre participação para embasar sua tese. Dentre as teorias estão: o Aprendizado Social, que refere-se àquilo que um usuário vê outros fazendo; *Feedback*, que é relativo aos efeitos que os usuários já presentes possuem sobre os novos; e Distribuição, relacionada à estrutura do conteúdo e ao alcance dele por meio da participação. Já as hipóteses são: Aprendizado Social - novos participante cujos amigos compartilham bastante conteúdo também irão compartilhar bastante; Marcação - novos participantes que são apontados em conteúdo irão contribuir mais com conteúdo; *Feedback* - novos participantes que recebem mais *Feedback* nos primeiros conteúdos que publicam irão compartilhar mais conteúdo; Distribuição - novos participantes cujo conteúdo se espalhou de forma significativa irão compartilhar mais conteúdo. Assim, foram feitas análises com os conteúdos agregados de 140.292 pessoas e também entrevistas onde os colaboradores da rede puderam demonstrar como utilizavam a ferramenta. Não foram feitas perguntas diretas sobre o que os motivou a compartilhar textos, fotos, comentários. Com isso autores conseguiram fundamentar as hipóteses 1, 3 e 4, e parcialmente 2. Neste último caso, explicam que este atributo não representou destaque significativo em relação à amostra utilizada. Segundo eles, pode-se dizer que o fato de ser marcado por um colega em uma foto só fez diferença para aqueles usuários que ainda não estavam muito familiarizados com a rede social, e muitos participantes da pesquisa demonstraram não compreender muito bem a funcionalidade.

O artigo de Wasko e Faraj (2005), por sua vez, aborda teorias de ação coletiva com a finalidade de compreender como as motivações pessoais e o capital social influenciam o compartilhamento de conhecimento nas redes virtuais. Os autores concluíram que as pessoas tendem a partilhar seu conhecimento sem esperar reciprocidade e fazem isso em três situações: quando entendem que isso realça sua reputação profissional; quando têm uma experiência que merece ser compartilhada; e quando se sentem incorporados na rede.

Além destes, outros trabalhos também abordam as motivações de usuários para suas contribuições nas redes sociais. Segundo Lindqvist et al. (2011), alguns aplicativos podem oferecer cupons de desconto com base na localização, enquanto aspectos de gamificação também são importantes aliados na hora de motivar pessoas a utilizarem um

serviço. Em seu estudo, Pelechrinis e Krishnamurthy (2012) descobriram que as pessoas passam a utilizar um sistema com mais frequência conforme ele se torna mais familiar para elas.

Junto a tais estudos, o de Santos et al. (2017) analisou os efeitos dos mecanismos de incentivo do Foursquare. Os mecanismos que o Foursquare apresenta são: Prefeitura e *Badges*. Um usuário se torna prefeito de um local quando é responsável pela maior quantidade de *check-ins* em um local nos últimos 30 dias. Os *badges* são concedidos de acordo com a localização, frequência de *check-ins*, eventos específicos ou datas comemorativas dos usuários. Foram analisados os perfis de 901 usuários a partir do dia que criaram seus perfis, por 13 semanas. Os perfis foram separados em grupos: grupo1 - usuários que fizeram até 250 *check-ins* nesse período (conservadores e talvez não tão motivados); grupo2 - usuários que fizeram entre 250 e 500 *check-ins* nesse período (um pouco mais motivados); grupo3 - usuários que fizeram mais de 500 *check-ins* nesse período (usuários constantes do sistema); Concluiu-se que alguns selos como *"Newbie"* e *"4sqDay2012"* não tem influência na adesão ao Foursquare, uma vez que são dados ao usuários logo no início e, portanto, aparecem muitas vezes. Já *badges* como *"Adventurer"*, *"SuperUser"*, *"Local"*, e *"Explorer"*, começam a ser mais frequentes entre usuários do grupo2 e são os mais comuns entre usuários do grupo3, o que faz com que provavelmente estejam associados com a motivação dos usuários. Já no que diz respeito às Prefeituras, estas tem efeito direto sobre os *check-ins*, uma vez que para se manter como prefeito é preciso continuar fazendo *check-ins*.

Como foi demonstrado, há diferentes aspectos que podem ser considerados como base para o desenvolvimento de tecnologias que auxiliem as atividades turísticas, sejam elas a escolha de destinos ou definição de rotas. Justamente no que diz respeito a tais aspectos é que esta pesquisa, de caráter exploratório, se diferencia das demais supracitadas. Outros estudos foram desenvolvidos a partir da análise de atributos demográficos ou de locais previamente visitados pelos usuários. De forma diferente, o trabalho aqui apresentado tem como premissa a ideia de que as informações fornecidas pelas *tags* escolhidas e selos obtidos pelos usuários, que são atributos particulares do TripAdvisor, podem colaborar para o entendimento dos perfis dos colaboradores. O conhecimento das características dos usuários permite que essa informação seja utilizada para as mais diversas finalidades, como já demonstrado na seção de justificativa. Sendo assim, este trabalho poderá, então, vir a servir como referência para estudos futuros relacionados a este tema e a esta ferramenta em específico.

3 CONTEXTUALIZAÇÃO

Este capítulo aborda os conceitos que serviram de base para a realização deste trabalho. Primeiramente é apresentado o TripAdvisor, seu sistema de recompensas e os tipos de informações que são inseridas pelos usuários. Em sequência, são explicados os conceitos de *Web Crawler*, banco de dados não-relacional, NoSQL e MongoDB, bem como de mineração de dados.

3.1 TRIPADVISOR

De acordo com pesquisas realizadas em maio de 2016 pela comScore, companhia responsável por medições de comportamento de consumidores e marcas ao redor do planeta, o TripAdvisor é considerado o maior site de viagens do mundo (O'CONNOR, 2008). O TripAdvisor permite que os usuários pesquisem a respeito de voos, hotéis, restaurantes, aluguéis para temporada e atrações a serem visitadas. Apesar de fornecer informações de preços de voos e diárias de hotéis, por exemplo, o foco do site não é a comercialização de serviços. Diferentemente, o TripAdvisor funciona como uma rede social de viagens, em que todos os usuários podem colaborar com revisões e avaliações dos lugares onde estiveram.

Para participar como colaborador do TripAdvisor, o primeiro passo é se cadastrar. Feito isso, para cada contribuição do usuário a rede social oferece recompensas diferentes. O processo de cadastro bem como o sistema de recompensas serão tópicos discutidos nas seções a seguir.

3.1.1 INFORMAÇÕES DOS USUÁRIOS

O perfil do usuário traz informações sobre os colaboradores do site que vão muito além das avaliações feitas por eles. Ao se cadastrar, o usuário é convidado a preencher campos referentes a informações demográficas, tais como idade, sexo e localização, além de uma breve descrição a seu respeito (figura 1).

Figura 1: Tela Editar Perfil

Os campos não são de preenchimento obrigatório, e, quando preenchidos, as informações inseridas são posteriormente apresentadas no canto superior esquerdo da tela de perfil, conforme figura 2. Além das informações demográficas, os usuários podem, ainda, selecionar etiquetas de estilo de viagem, chamadas de *tags*. Estas etiquetas definem o tipo de viagem preferido pelo usuário. São dezenove opções diferentes e não há número mínimo nem limite para escolha, é possível definir quantas *tags* desejar. A coleção de *tags* disponibilizada pelo TripAdvisor é mostrada na figura 3.

As informações obtidas a partir das *tags* e de selos exercem um papel de extrema relevância no momento de se identificar perfis dos colaboradores da rede e, por esta razão, são os objetos principais de estudo deste trabalho.

3.1.2 SISTEMA DE RECOMPENSAS

Como forma de reconhecimento, o TripAdvisor possui um sistema que premia a contribuição dos usuários ao site, o TripColaboradores (TRIPADVISOR, 2016c). As premiações são feitas em forma de pontos, níveis e selos. As informações referentes à quantidade de pontos, bem como à coleção de selos, podem ser encontradas na página de perfil do usuário. Cada tipo de contribuição (Ex: revisão, foto, vídeo, etc.) gera uma pontuação diferente. A figura 4 mostra todas as possibilidades de contribuição e suas respectivas pontuações.

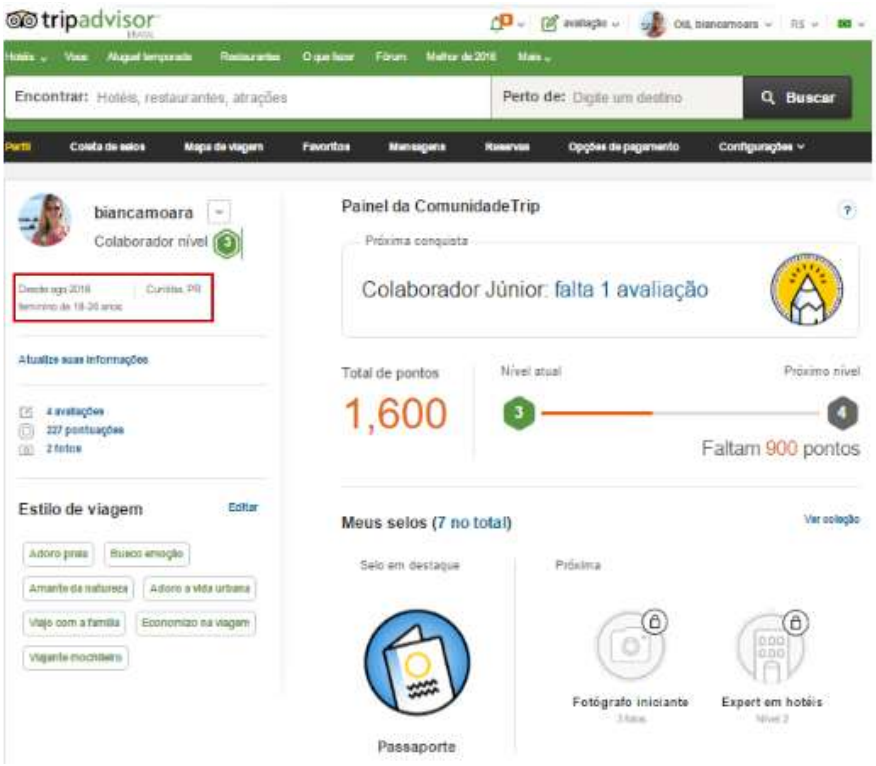

Figura 2: Tela Exemplo de Página de Perfil

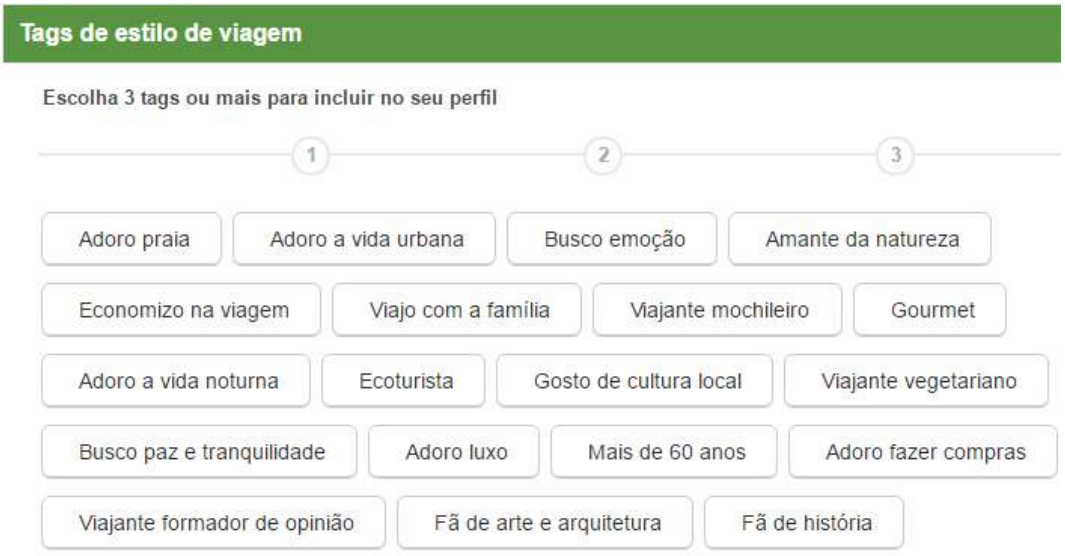

Figura 3: Tela Lista de Tags

A quantidade de pontos, por sua vez, define o nível em que o colaborador se encontra. Existem seis diferentes níveis que podem ser alcançados, conforme mostrado na figura 5. Quanto mais um usuário colabora, mais pontos ganha e, ao atingir o valor máximo correspondente a um nível, passa para o próximo.

Figura 4: Pontuações

Figura 5: Níveis

Os selos, segundo o TripAdvisor, são "uma forma de [o colaborador] exibir seu conhecimento e experiência" (TRIPADVISOR, 2016a). Existem diferentes categorias de selos e diferentes formas de obtê-los. São seis categorias de selos – Colaborador Iniciante, Especialista, Passaporte, Votos Úteis, Explorador e Colaborador do *Travellers' Choice* – e cada uma funciona de uma forma específica. Algumas categorias, como a de Especialista, possuem níveis. Entretanto, os níveis relativos aos selos são classificações válidas

apenas para selos, e nada tem a ver com o nível atingido com base na pontuação do colaborador. Em outras palavras, um usuário pode ser um colaborador de nível 3 (dentre os seis existentes) e, ao mesmo tempo, possuir um selo de *Expert* em hotéis de nível 1. Isso porque o nível de colaborador sobe de acordo com o número de pontos e o nível do selo de *Expert* em hotéis sobe de acordo com o número de avaliações sobre hotéis. De certa forma, pode-se dizer que os selos são formas de reconhecimento de ações específicas. A seguir são descritas mais detalhadamente as categorias de selos, seu funcionamento e como são atingidas.

- Selos Colaborador Iniciante

A categoria de selos Colaborador Iniciante é relativa à quantidade de avaliações realizadas pelos usuários. A cada determinado número de avaliações, conforme figura 6, o colaborador recebe um novo selo. Por exemplo, ao realizar uma avaliação, o usuário recebe o selo "Novo Colaborador". Para receber o próximo, selo "Colaborador", é preciso realizar mais duas avaliações, totalizando as três necessárias para tal nível, e assim por diante. É importante ressaltar que nesta categoria são adicionados novos selos de acordo com a evolução do colaborador, enquanto na categoria Especialista, por exemplo, que será descrita em detalhes a seguir, apenas atualiza-se o nível dos selos.

Figura 6: Exemplo de Selos Colaborador Iniciante

- Selos Especialista

No TripAdvisor, além de receberem selos pelo número de revisões que realizam, os usuários podem também receber selos de especialistas. Essa categoria possui três subcategorias em que é possível se tornar um *expert*. São elas: Especialista em hotéis, Especialista em restaurantes e Especialista em atrações. Em relação a hotéis, as subcategorias são ainda mais específicas. Além de Especialista em hotéis, o TripAdvisor possui selos para Especialista em hotéis luxuosos, Especialista em hotéis-boutique, Especialista em pousadas e Especialista em resorts. Nesta categoria os selos possuem níveis. Cada nível é atingido a cada três revisões sobre uma mesma subcategoria.

- Selo Passaporte

O selo passaporte é recebido quando o usuário realiza pelo menos duas avaliações em cidades diferentes. Este selo não possui níveis e é atualizado apenas por meio da atualização da lista de cidades avaliadas.

- Selos Impacto

Os selos de impacto são concedidos de acordo com o número de leitores que um colaborador possui. Essa categoria não possui níveis, o usuário recebe um novo selo a cada determinada quantidade de leitores que atinge.

- Selos Votos Úteis

Este selo é concedido a partir do reconhecimento da comunidade do TripAdvisor. Se a revisão feita por um usuário for considerada útil por outro colaborador, este pode conceder um voto útil ao usuário que fez a avaliação. Por esse motivo, os selos de votos úteis podem ser considerados uma forma de reconhecimento à credibilidade de uma avaliação (AMARAL et al., 2014). Nesta categoria, o selo de Colaborador Importante é atualizado de acordo com o número de votos úteis recebidos, como pode ser visto na figura 7.

- Selo Explorador

O selo Explorador é concedido quando o usuário faz uma das primeiras avaliações de um hotel, restaurante ou atração em um determinado idioma. Essa categoria não possui níveis, apenas são listados os lugares avaliados, vide figura 8.

Figura 7: Selos Votos Úteis

Figura 8: Selo explorador

- Selo Colaborador do *Travellers' Choice*

Travellers' Choice, figura 9, é um prêmio concedido a hotéis, restaurantes e atrações considerados os melhores do mundo de acordo com revisões feitas por milhões de viajantes. Este selo é conquistado quando o colaborador avalia de maneira positiva um local que tenha sido premiado (TRIPADVISOR, 2016b).

Figura 9: Selo Colaborador do *Travellers' Choice*

3.2 *WEB CRAWLERS*

Segundo Benevenuto et al. (2012), acessar dados armazenados em servidores de aplicações de redes sociais online não é tarefa fácil. Como mostram, são pouquíssimos os trabalhos que utilizam dados obtidos diretamente destes servidores (BALUJA et al., 2008), (CHUN et al., 2008), (DUARTE et al., 2007). Por este motivo, uma estratégia para coleta de dados destas bases é utilizar os chamados robôs ou *web crawlers*.

Um *web crawler* é um agente de software ou um tipo de robô de Internet automatizado sistematicamente para realizar tarefas de modo recursivo em páginas *web* (CHEONG, 1996). Sites de busca na Internet fazem uso de *web crawlers* para fornecer dados atualizados, além de prover consultas mais rápidas. Isso se dá pelo fato desse tipo de agente ser utilizado, em sua maioria, para realizar cópias de sites, podendo assim indexá-los, fazendo com que as ferramentas de busca tenham melhores tempos de resposta das consultas realizadas. Alguns dos *crawlers* mais conhecidos são:

- Googlebot - *crawler* do Google.
- Yahoo! Slurp - *crawler* do Yahoo!.
- Msnbot - *crawler* do Bing (Microsoft).

Atualmente, existem *web crawlers* para *web* semântica, que possuem dados representados por um padrão e seguem alguns princípios, como tornar as informações mais legíveis e atribuir algum significado ou sentido para a palavra informada pelos usuários (MACHADO et al., 2016). Existem também *crawlers* que extraem informações específicas de sites para serem utilizadas no futuro, como buscadores de *e-mail* para envio de *spams*.

O comportamento de um *crawler* na *web* pode ser definido por uma combinação de políticas, conforme Castillo (2005) define:

- Política de revisitação: quando o programa deve checar atualizações nas páginas;
- Política de seleção: quais conteúdos ou páginas serão extraídos;
- Política de paralelismo: como coordenar agentes distribuídos;
- Política de boas maneiras: como evitar a sobrecarga de *websites*.

Respeitando as regras definidas é possível desenvolver um *web crawler* que terá um bom desempenho computacional sem prejudicar a origem da informação extraída. Alguns tipos de coletas realizadas em redes sociais online são: utilização de APIs, coleta por amostragem, coleta em larga escala, coleta por inspeção de identificadores e coleta em tempo real.

Uma API é "um conjunto de tipos de requisições HTTP juntamente com suas respectivas definições de resposta"(BENEVENUTO et al., 2012). Apesar de serem excelentes para a coleta de dados de redes sociais online, uma vez que oferecem os dados em formatos estruturados (XML e JSON), nem todos os sistemas possuem APIs, ou, então, a API não está disponível para ser utilizada com a finalidade necessária. Este último caso é o do TripAdvisor. Sua API é disponibilizada apenas para *websites* e aplicativos de viagens voltados para o consumidor (TRIPADVISOR, 2018). Uma vez que este estudo não se encaixa neste perfil, optou-se então pela criação e uso de *web crawlers*.

A coleta por amostragem é utilizada em casos em que não é possível coletar o grafo inteiro de uma rede social online, apenas parte dele. Algumas estratégias de coleta por amostragem são: *snowball* (ou bola de neve), coleta do maior componente fracamente conectado (WCC), e coleta baseada em caminhadas aleatórias (ou *random walks*). De forma resumida, a estratégia *snowball* consiste em realizar uma busca em largura para coletar o grafo de uma rede social online. A coleta é iniciada a partir de um nodo semente e, quando a lista de vizinhos desse nodo é coletada, novos nodos são descobertos e coletados no segundo passo. Este termina apenas quando todos os nodos descobertos no primeiro passo são coletados. No próximo passo todos os nodos descobertos anteriormente são coletados, e assim sucessivamente. Idealmente deve-se utilizar uma grande quantidade de nodos sementes a fim de evitar que a coleta seja restrita apenas a um pequeno componente do grafo. A coleta do maior componente fracamente conectado (WCC) é considerada interessante uma vez que o maior WCC de um grafo é, de forma estrutural, a parte

mais interessante de ser analisada. Isso porque é o componente que registra a maior parte das atividades dos usuários (MISLOVE et al., 2007). Uma das opções de coleta baseada em caminhadas aleatórias parte de um nodo semente v e prossegue selecionando aleatoriamente um dos vizinhos de v sucessivamente até que um número B pré-definido de nodos tenha sido selecionado. Nesta abordagem um mesmo nodo pode ser selecionado múltiplas vezes (BENEVENUTO et al., 2012).

A coleta em larga escala, utilizada como forma de coleta de grandes bases de dados de redes sociais online, faz uso de coletores distribuídos em diversas máquinas. Isso se dá a fim de evitar que a coleta de dados públicos seja interpretada como um ataque aos servidores de redes sociais e também por causa do processamento necessário para tratar e salvar os dados coletados. Uma das estratégias para realizar tal coleta consiste em utilizar uma máquina mestre e máquinas escravas. A máquina mestre possui uma lista centralizada de usuários a serem visitados, enquanto as máquinas escravas coletam, armazenam e processam os dados coletados com a finalidade de identificar novos usuários. Uma vez identificados, novos usuários são repassados para a máquina mestre, que os distribui para as máquinas escravas (BENEVENUTO et al., 2012).

A coleta por inspeção de indicadores é viável em sistemas que atribuem um identificador (ou ID) numérico e sequencial para cada um dos usuários cadastrados. Neste caso é possível realizar uma coleta da rede completa, ao invés de apenas parte dela. Uma vez que novos usuários recebem um identificador sequencial, torna-se possível percorrer todos os IDs, sem a necessidade de verificar a lista de amigos desses usuários em busca de novos IDs. Exemplos deste tipo de sistema são o MySpace e o Twitter (BENEVENUTO et al., 2012).

Por fim, a coleta em tempo real, como o próprio nome sugere, é referente à coleta de informações propagadas por usuários em tempo real. Tais informações tem sido usadas para diferentes finalidades, como consultas a informações temporais e geográficas, por diversas aplicações, como por exemplo *Google Insights* e Bing (BENEVENUTO et al., 2012).

3.3 BANCO DE DADOS NÃO-RELACIONAL (NOSQL)

Bancos de dados não-relacionais são bancos que foram desenvolvidos para gerenciar dados com esquema flexível, volumoso e heterogêneo, a fim de garantir escalabilidade (CROCKFORD, 2006). Esse tipo de banco não utiliza o esquema tabular de linhas e

colunas. Ao contrário dos bancos de dados mais tradicionais, estes possuem um modelo de armazenamento que é otimizado de acordo com os requisitos específicos do tipo de dado que está sendo registrado (WASSON; TEJADA, 2017).

O termo NoSQL refere-se a repositórios de dados que, ao invés de SQL, usam outras linguagens de programação e construções para consultar os dados. Na prática, NoSQL significa "banco de dados não-relacional", apesar de muitos desses bancos oferecerem suporte a consultas compatíveis com SQL. No entanto, a estratégia de execução de consultas é geralmente muito diferente da maneira como um Sistema Gerenciador de Banco de Dados (SGBD) tradicional executaria a mesma consulta SQL (WASSON; TEJADA, 2017).

Como dito, os bancos não-relacionais armazenam os dados de maneira diferente dependendo dos requisitos apresentados por eles. Algumas categorias de bancos NoSQL são as que armazenam: dados em documentos, dados em colunas, dados gráficos, objetos e índices externos. No caso deste trabalho, utilizou-se o armazenamento em documentos. Uma vez que são variados os tipos de informações coletadas dos perfis dos usuários (*badges*, *tags*, texto), julgou-se esta ser uma boa forma de registrar as informações de seus perfis. O banco NoSQL escolhido foi o MongoDB.

O MongoDB é um banco de dados da categoria dos NoSQL. O banco de dados implementa o modelo de dados orientado a documento, ou seja, os dados são organizados na forma de documentos que representam conjuntos de chave-valor. Fazendo associação com o modelo relacional, um documento representa uma tupla e uma coleção, ou seja, um conjunto de documentos seria compreendido como uma tabela (WALTER; DUARTE, 2016). Um documento é apresentado utilizando o formato JSON (CROCKFORD, 2006).

O MongoDB possui muitas das características dos bancos de dados relacionais, tais como índices secundários e consultas dinâmicas, além de alta capacidade para dimensionar. Junto a isso, manter os dados em documentos pode fazer com que as consultas sejam mais rápidas do que em um banco de dados relacional, onde os dados ficam separados em várias tabelas (PESSOA et al., 2012).

A utilização do MongoDB é de grande importância no contexto da *web*, onde a carga de informações pode aumentar de repente, pois com o MongoDB é possível aumentar sua capacidade sem qualquer tempo de inatividade (PESSOA et al., 2012).

3.4 MINERAÇÃO DE DADOS

3.4.1 O QUE É MINERAÇÃO DE DADOS?

No livro *Data Mining: Concepts and Techniques* os autores fazem uma analogia com a mineração de ouro e pedras preciosas para ajudar a definir o conceito de mineração de dados. Segundo eles, quando se trata de minerar o ouro a partir de pedras comuns e areia, não dizemos "mineração de pedra" ou "mineração de areia", e sim, "mineração de ouro". Da mesma forma, portanto, os autores acreditam que o termo "mineração de dados" não traz a correta definição do processo proposto, sendo mais adequado dizer "mineração do conhecimento". Isso porque, assim como na extração de minérios, neste caso o que se busca não são apenas dados, mas, sim, obter conhecimento relevante a partir deles. Sendo assim, em outras palavras, a mineração de dados pode ser definida como o processo de descoberta de padrões e conhecimento a partir de grandes quantidades de dados. O termo, talvez por ser mais curto, pode ser muitas vezes interpretado como o processo de descoberta de conhecimento a partir de dados (em inglês, *Knowledge Discovery from Data – KDD*).

Entretanto, a mineração de dados pode também ser considerada apenas uma parte deste processo. De acordo com Han et al. (2012), o processo de KDD envolve os seguintes passos:

- Limpeza de dados (*Data cleaning*)
- Integração de dados (*Data integration*)
- Seleção de dados (*Data selection*)
- Transformação de dados (*Data transformation*)
- Mineração de dados (*Data mining*)
- Análise de padrões (*Pattern evaluation*)
- Apresentação do conhecimento (*Knowledge presentation*)

Nessa definição, os quatro primeiros passos têm como função preparar os dados brutos para serem minerados. Sendo assim, a primeira fase, de limpeza de dados, como o próprio termo sugere, refere-se à remoção de ruídos e dados inconsistentes do conjunto de dados; a segunda, integração de dados, serve como forma de combinar dados provenientes

de diferentes fontes, que são armazenados em *data warehouses*; a terceira, seleção de dados, é a etapa em que os dados mais relevantes são selecionados para análise; e a quarta, transformação de dados, onde os dados são transformados e consolidados em formatos mais apropriados para serem minerados. Após o estágio de mineração, quando são descobertos padrões nos dados em estudo, acontecem as etapas de análise e possível apresentação de tais descobertas.

Seja como for, a mineração de dados é essencial quando se trata de analisar grandes volumes de dados. É a partir dela que padrões inicialmente não reconhecíveis passam a ser facilmente percebidos e podem, então, ser analisados.

3.4.2 PARA QUE UTILIZAR MINERAÇÃO DE DADOS?

A Internet já é parte integrante da vida de mais da metade da população mundial. Isso pode ser constatado a partir de informações geradas pelo estudo do *We Are Social* e *Hootsuite*, *Digital in 2017 Global Overview*, com conteúdo sobre o Brasil e o mundo. O estudo mostra que mais da metade da população mundial utiliza um *smartphone* e mais da metade do tráfego de informações da web mundial vem de aparelhos de telefonia móvel. Em relação ao Brasil, 139 milhões de brasileiros utilizam a Internet, o que equivale a 67% da população (MARTINS, 2017). Além disso, metade dos brasileiros acessa a Internet a partir de dispositivos móveis, 58% utiliza redes sociais e 90% faz uso da Internet todos os dias. Gastamos, por dia, em média 8h56min conectados (MARTINS, 2017).

Tanta conectividade gera um enorme volume de dados. Segundo estudo divulgado pela Business Software Alliance (BSA) em 2015, 2.5 quintilhões de *bytes* (equivalente a 2.5 bilhões de *gigabytes*) são criados todos os dias (CIO, 2015). Pesquisas mostram que a tendência, até 2020, é termos disponíveis 40 trilhões de *gigabytes* de dados (ALBUQUERQUE, 2017). Esse crescimento expressivo do volume de dados disponível é resultado da informatização da sociedade, vide dados apresentados no parágrafo anterior, e do rápido desenvolvimento de ferramentas de coleta e armazenamento de dados (ALBUQUERQUE, 2017). Foi neste contexto, da necessidade de obter informações importantes a partir de uma vasta quantidade de dados e transformar tais dados em conhecimento, que surgiu a mineração de dados.

3.4.3 FERRAMENTAS PARA MINERAÇÃO DE DADOS

Existem diferentes ferramentas de mineração de dados. Estas são utilizadas para definir padrões em determinadas tarefas que podem ser descritivas ou preditivas. Tarefas descritivas são aquelas para as quais se utiliza uma ferramenta de mineração a fim de descobrir características, propriedades dos dados disponíveis em um *dataset*. Já em uma tarefa preditiva utiliza-se uma ferramenta de mineração com a intenção de realizar previsões acerca dos dados disponíveis. De acordo com Han et al. (2012), exemplos de ferramentas são: caracterização e discriminação; mineração de padrões frequentes, associações e correlações; classificação e regressão; análise de agrupamentos (*clusters*); e análise anormal (*outlier analysis*). A ferramenta utilizada neste trabalho foi a mineração de padrões frequentes, associações e correlações.

Padrões frequentes referem-se a padrões que ocorrem com frequência quando se trata de dados sendo analisados. Dentre os possíveis padrões frequentes estão os *itemsets*. Um *itemset* frequente é um conjunto de itens que aparecem juntos com frequência em um conjunto de dados transacional [1] (HAN et al., 2012). A mineração de padrões frequentes leva à descoberta de associações e correlações entre os dados deste conjunto. Uma regra de associação só é considerada válida quando satisfaz limiares de suporte e confiança mínimos (*thresholds*). O suporte corresponde ao número de transações que contém o conjunto de itens em questão (VASCONCELOS; CARVALHO, 2004). A confiança corresponde ao grau de certeza da regra encontrada (HAN et al., 2012).

Um exemplo de aplicação desse conceito seria buscar saber com que frequência os clientes de uma rede de supermercado, ao comprarem o produto A, compram também o produto B. Mas para quê? Um famoso caso é o da relação entre fraldas e cervejas. Inicialmente esses itens parecem completamente desconexos. No entanto, em um supermercado americano foi descoberto que havia uma ligação entre a venda destes produtos, uma vez que homens, quando saíam para comprar fraldas, aproveitavam para levar cervejas. A partir de tal constatação houve uma mudança na disposição dos produtos no mercado – cervejas foram colocadas ao lado das fraldas – e, então, um aumento considerável nas vendas (GUROVITZ, 2011). Este exemplo mostra que a utilização de regras de associação permite identificar padrões e perfis de consumidores, e, consequentemente, guiar ações mais focadas em um determinado objetivo, seja o aumento de lucratividade, retenção de clientes, sugestões de produtos e/ou serviços, entre outros. Neste trabalho cada transação

[1] Uma transação corresponde a um conjunto formado por um número de identificação único e uma lista com os itens que a compõem.

contém o ID do usuário e suas ações no TripAdvisor, como as cidades visitadas, as *tags* marcadas e os selos colecionados.

Em termos formais, uma regra de associação seria definida da seguinte forma (AGRAWAL; SRIKANT, 1994):

> Consideremos L = {i1, i2, i3, ..., in} um conjunto de literais, chamado de itens, e D um conjunto de transações, onde cada transação T é um conjunto de itens tal que $T \subseteq L$. Associado a cada transação está um identificador único, chamado TID. É dito que uma transação T contém X, um subconjunto de itens de L, se $X \subseteq T$. Uma regra de associação, portanto, é uma implicação da forma $X \Rightarrow Y$, onde $X \subset L$, $Y \subset L$, e $X \cap Y = \phi$;. A regra $X \Rightarrow Y$ para um conjunto de transações D possui confiança c se c% das transações em D que contém X também contém Y. A regra $X \Rightarrow Y$ tem suporte s no conjunto de transações D se s% das transações em D contém $X \cup Y$. Sendo assim, dado um conjunto de transações D, o problema de minerar regras de associação é referente a gerar todas as regras de associação que possuam suporte e confiança maiores do que os mínimos pré-estabelecidos por usuários, chamados minsup e minconf, respectivamente. Neste contexto, D pode ser um arquivo, uma tabela ou o resultado de uma expressão relacional, por exemplo.

Para mineração de *itemsets* frequentes deste trabalho utilizou-se o algoritmo *apriori*. Este será descrito em detalhes na seção 4.

4 METODOLOGIA

Para melhor descrever a metodologia utilizada no desenvolvimento deste trabalho, a mesma foi dividida nas seguintes fases: escolha das ferramentas; Desenvolvimento de *web crawlers* e extração dos dados do TripAdvisor; armazenamento dos dados obtidos em um banco de dados não-relacional; limpeza dos dados obtidos; mineração dos dados; e análise dos dados.

Estas fases serão melhor descritas nas seções ao longo deste capítulo. O processo metodológico pode ser representado pela figura 10.

Figura 10: Passos metodológicos

4.1 ESCOLHA DAS FERRAMENTAS

O TripAdvisor foi escolhido como objeto de estudo, como já citado anteriormente na seção de justificativa, por apresentar recursos específicos, os *badges* e as *tags* de estilo e viagem. Apostou-se que a análise destes atributos pode contribuir para atingir o objetivo deste trabalho, de compreender o comportamento dos usuários a partir de informações presentes nos seus perfis.

Para tal, foi necessário extrair as informações dos perfis dos usuários. Primeira-

mente, procurou-se utilizar a API do próprio TripAdvisor. Porém, como informado pela página, "A API do conteúdo do TripAdvisor é apenas para sites e aplicativos de viagens voltados para o consumidor"(TRIPADVISOR, 2018). Uma vez que nosso objetivo era de pesquisa, e a quantidade de dados disponibilizada pela API não seria o suficiente, optou-se por utilizar *web crawlers*.

Após a extração, os dados precisariam ser armazenados em um banco de dados. Foi escolhido um banco não relacional, o MongoDB, por este ter sido desenvolvido com a finalidade de gerenciar dados cujo formato pode ser variável e que são também muito volumosos, que é o caso dos dados coletados a partir do TripAdvisor. Para uma melhor visualização dos dados, foi instalada a ferramenta RoboMongo, que faz a conexão com a base de dados criada e serve como interface gráfica para consultas.

Uma vez armazenados, foi aplicado sobre os dados o algoritmo *apriori* a fim de extrair regras de associação entre eles. Para obtenção e auxílio na análise das regras utilizou-se o RStudio.

4.2 DESENVOLVIMENTO DE *WEB CRAWLERS* E EXTRAÇÃO DOS DADOS DO TRIPADVISOR

Para realizar a extração dos dados de análise, utilizou-se o próprio site e suas características para criar um sistema com dois *Web Crawlers* capazes de identificar os usuários e gravar suas informações em uma base de dados.

O primeiro começa sua busca pela URL de uma cidade presente no TripAdvisor (no caso deste trabalho, a cidade escolhida foi Curitiba). Cada cidade, por sua vez, tem uma lista de hotéis, atrações e restaurantes que podem ser visitados e avaliados pelos usuários. Quando o usuário deixa sua avaliação, é possível identificar a URL da sua página de perfil, que contém as informações a serem extraídas. Em resumo, o primeiro *web crawler* busca iterar recursivamente entre os links e páginas onde existem avaliações dos usuários, retirando os links de perfis para análise. Essa lista é gravada em um arquivo de texto que serve de entrada para o segundo *web crawler*.

O segundo *web crawler*, por sua vez, faz a varredura dos links, visitando cada perfil e selecionando todas as informações disponíveis. Tendo as informações selecionadas, estas são então armazenadas em um banco de dados não-relacional.

4.3 ARMAZENAMENTO DOS DADOS OBTIDOS

O formato das informações pode variar de usuário para usuário, visto que um usuário pode ser mais ativo e ter mais recompensas alcançadas que outro, ou ter um perfil mais detalhado. Logo, gravar os dados em tabelas comuns de bancos de dados relacionais se torna uma tarefa difícil, não sendo possível prever quantos atributos uma entidade pode ter. Para resolver este problema, foi utilizado um banco de dados orientado a documentos, o MongoDB. Ao coletar os dados de um usuário, o *web crawler* faz uma conversão do seu objeto de dados para a extensão JSON, que pode ser inserida diretamente no contexto criado dentro da base de dados. Após a inserção dos dados de cada usuário, a base está pré-definida para realização de análises de padrões e associação.

4.4 LIMPEZA DOS DADOS OBTIDOS

Depois de armazenados os dados no banco de dados, foram extraídos dois documentos em formato .txt e um em formato .csv. Dos documentos com extensão .txt, um continha as *tags* informadas pelos usuários e outro as informações sobre as cidades visitadas por eles. Já o documento .csv apresentava as informações sobre os *badges* colecionados. O arquivo com as cidades continha também os *ids* dos usuários. Este e o de *badges* eram muito extensos e, por esse motivo, foi necessário que passassem por um processo de limpeza de dados antes de serem analisados. O processo de limpeza de dados para cada um dos arquivos será explicado nas subseções seguintes.

4.4.1 LIMPEZA DOS DADOS DAS CIDADES VISITADAS

No processo de limpeza do arquivo de cidades foram removidos: URL das páginas de perfil, números, *underline*, traços, vírgulas e espaços. Acentos e cedilha foram substituídos pelas respectivas letras sem sinal de acentuação. A ordem seguida de remoção e substituição foi: remoção das URLs das páginas de perfil, remoção de números, substituição de acentos, remoção de *Underline* e traços, substituição de cedilha, remoção de vírgulas, remoção de espaços.

O resultado desta limpeza inicial foi um documento apenas com nomes das cidades, separadas por vírgulas. Esse procedimento foi necessário para, no momento da análise, não haver erros de contabilização. Alguns *ids* de usuários, como por exemplo, “fabiCuritiba”podiam ter parte do texto confundido com uma aparição da cidade Curitiba na hora de utilizar uma ferramenta para fazer a contagem das cidades. Outro caso

também aconteceria se os espaços não fossem removidos, pois a cidade de Nova Londrina, por exemplo, poderia ter parte do nome contabilizado junto com as aparições da cidade "Londrina".

O passo seguinte foi fazer a contagem das cidades. Para tal, utilizou-se a ferramenta online *Word Frequency Counter* (WRITEWORDS, 2018), que conta a quantidade de registros de cada palavra. Feito isso, os dados ficaram prontos para serem minerados.

4.4.2 LIMPEZA DOS DADOS DE *BADGES*

A limpeza de dados referente aos *badges* ocorreu da seguinte maneira: primeiramente comparou-se a lista de selos disponíveis pelo TripAdvisor com a lista dos selos da base de dados coletada. Então, dentre os *badges* de maior ocorrência na base, foram escolhidos aqueles que julgou-se que teriam mais impacto na identificação de perfis dos usuários.

Neste sentido, foram desconsiderados, por exemplo, os *badges* de fotografia, já que nenhuma imagem teria papel significativo na descrição do perfil dos usuários. O selo de Novo Colaborador, dado a todos os usuários quando realizam a primeira avaliação, também não apresentou relevância, uma vez que é parte de todas as listas de *badges*.

Portanto, os *badges* levados em consideração foram aqueles julgados como mais relevantes no contexto deste estudo: os selos de especialidade (Expert em Hotéis, Expert em Restaurantes e Expert em Atrações) e Explorador.

4.5 MINERAÇÃO E ANÁLISE DOS DADOS

Como dito anteriormente na seção 3, o algoritmo utilizado neste trabalho para encontrar as regras de associação foi o *Apriori*. O *Apriori* se utiliza de duas funções: *apriori-gen*, para gerar o conjunto de itens candidatos, eliminando os que não são frequentes, e *gen-rules*, para extrair as regras de associação. O objetivo do algoritmo é buscar correlações entre os dados na forma de *itemsets* frequentes (que têm suporte igual ou maior ao suporte mínimo pré-estabelecido). Ao mesmo tempo, o *apriori* calcula os valores de confiança e suporte. É preciso fornecer 3 entradas ao algoritmo: um valor para o suporte mínimo, outro para confiança mínima e um arquivo de itens e transações (VASCONCELOS; CARVALHO, 2004). A execução do algoritmo acontece da seguinte maneira:

- Primeira iteração: é feita a contagem do suporte para cada conjunto de tamanho unitário (1-itemsets). Todos os conjuntos que satisfazem o suporte mínimo são selecionados.

- Segunda iteração: é feita a junção de conjuntos 1-itemset pela função apriori-gen e são gerados os conjuntos 2-itemsets. Os suportes são definidos por meio de pesquisa no banco de dados, e assim, são encontrados os conjuntos 2-itemsets frequentes.

- Demais iterações: o algoritmo prossegue iterativamente, até que o conjunto de k-itemsets encontrado seja um conjunto vazio.

Os dados sobre cidades, *badges* e *tags* presentes no banco de dados foram aplicados ao algoritmo, utilizando diferentes valores de confiança e suporte para alinhar a importância das regras geradas de acordo com o tamanho das informações.

5 RESULTADOS OBTIDOS

Neste estudo contou-se com uma base de dados de 42.381 usuários que fizeram pelo m enos uma revisão para a cidade de Curitiba. Do perfil de cada um deles foram extraídas informações referentes às *tags* escolhidas, aos *badges* conquistados e às cidades avaliadas. Este capítulo engloba todos os resultados obtidos a partir do desenvolvimento deste trabalho e está dividido em duas seções: Visão Geral dos Dados (seção 5.1) e Perfis dos Usuários - Aplicação do algoritmo *apriori* (seção 5.2), explicadas a seguir.

5.1 VISÃO GERAL DOS DADOS

Esta seção apresenta uma visão geral dos resultados obtidos para cada um dos tipos de informação coletados (cidades, *tags* e *badges*, respectivamente).

5.1.1 INFORMAÇÕES SOBRE CIDADES AVALIADAS OBTIDAS PELO SELO PASSAPORTE

Nesta seção são apresentados os resultados obtidos a partir dos conteúdos do selo passaporte. Como descrito anteriormente na seção 3, o selo passaporte é adquirido quando o colaborador realiza avaliações para pelo menos duas cidades diferentes.

O gráfico (figura 11) e a nuvem de palavras (figura 12) mostram para quais cidades os 42.381 colaboradores cujos perfis foram analisados neste estudo mais fizeram revisões. São apresentados os resultados para as 27 cidades mais visitadas. Isso porque os resultados tendem a estabilizar após este valor, não havendo diferença relevante a ser evidenciada.

A base de dados deste trabalho é referente às informações de usuários que avaliaram a cidade de Curitiba. Por esta razão, Curitiba é a cidade que apresenta mais resultados: 14.444 registros. Esta diferença em relação ao total coletado na base (42.381) pode ser notada uma vez que o selo Passaporte só é concedido quando o usuário avalia ao menos duas cidades. Sendo assim, na base há registros de colaboradores que fizeram

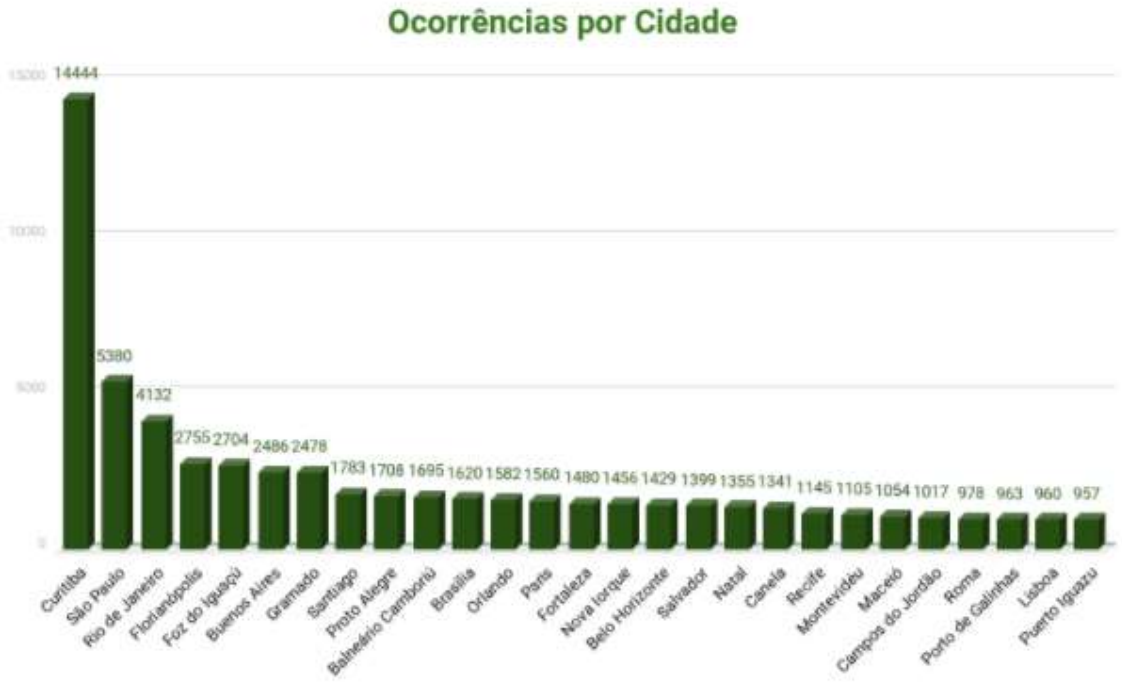

Figura 11: Gráfico do número de ocorrências de cada cidade

Figura 12: Nuvem de palavras referente às ocorrências de cada cidade

apenas uma revisão, para Curitiba, e por isso são contabilizados mas não receberam o selo.

As demais são cidades que também foram avaliadas por colaboradores que se

referiram à Curitiba no TripAdvisor. A segunda mais frequente é São Paulo, que, com uma diferença de 9.064 ocorrências para a primeira, foi registrada 5.380 vezes, seguida pelo Rio de Janeiro, com 4.132. A capital catarinense, Florianópolis, vem logo após, com 2.755 aparições, juntamente com Foz do Iguaçú (2.704), Buenos Aires (2.486) e Gramado (2.478). A oitava cidade a aparecer na lista é Santiago (1.783), e é a última cidade a apresentar diferença considerável com relação à cidade anterior, de 695 ocorrências. A partir daí os registros decrescem com uma variação menos abrupta. Entre as cidades de Canela (1.341 registros), no Rio Grande do Sul, e Recife (1.145 registros), em Pernambuco, a diferença é de somente 196 ocorrências. As últimas cidades mostradas no gráfico, Porto de Galinhas (963), Lisboa (960) e Puerto Iguazú (957) têm, entre si, diferença de apenas 3 ocorrências. Por este motivo, como já dito, a partir deste ponto os resultados não se mostraram tão significativos e, portanto, foi decidido considerar apenas essas 27 primeiras cidades.

Dentre as cidades selecionadas, podemos contar 18 cidades brasileiras, das quais 12 são capitais. O gráfico representado pela figura 13 traz as 27 cidades analisadas agrupadas por região. Nota-se que a maioria das cidades brasileiras pertence à região Sul (7.26%), seguidas pela região Nordeste (6.22%), Sudeste (4.15%) e Centro-Oeste (1.4%). Não houve registro para a região Norte. Já os destinos internacionais representam um terço das cidades selecionadas. Destas, 4 cidades pertencem à América do Sul (Buenos Aires, Santiago, Montevideu e Puerto Iguazú), 3 à Europa (Paris, Roma e Lisboa) e duas aos Estados Unidos (Nova Iorque e Orlando).

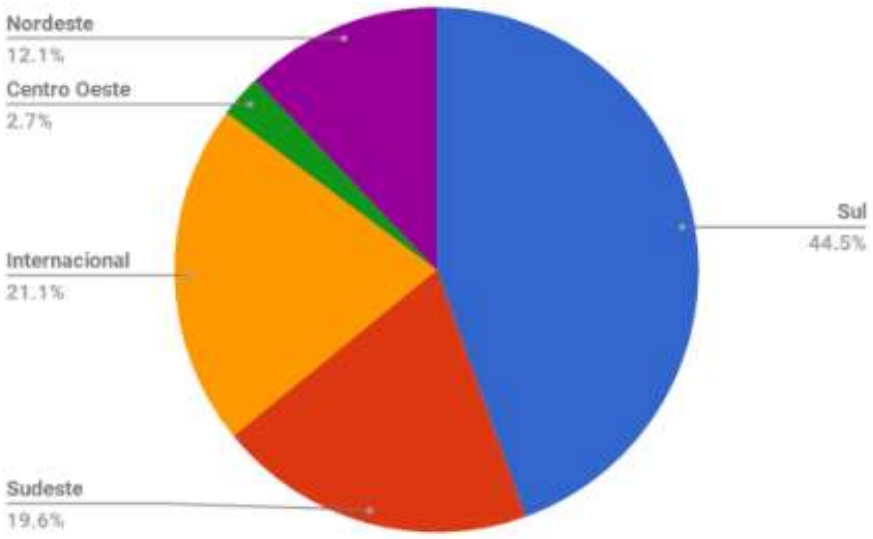

Figura 13: Gráfico das 27 cidades mais visitadas/avaliadas agrupadas por região

No entanto, se analisarmos o gráfico que compara o total de ocorrências somente para capitais e as agrupa por região (figura 14), veremos que as capitais com mais registros estão presentes na região Sul (18906 - 44%), seguida pelas regiões Sudeste (11544 - 27%), Nordeste (8278 - 19%), Centro-Oeste (3117 - 7%) e Norte (1401 - 3%). A figura 15 traz o mapa do Brasil. Nele os valores presentes representam o número de ocorrências para a capital de cada um dos estados e do Distrito Federal.

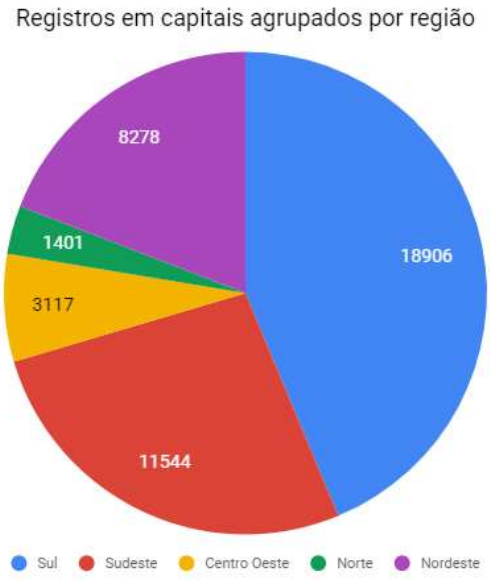

Figura 14: Gráfico do número de ocorrências de capitais por região

Os dados obtidos a partir do selo Passaporte ajudam a compreender as preferências dos usuários no que diz respeito às cidades e regiões que costumam visitar. Essas informações podem ser utilizadas, por exemplo, por agências de turismo e companhias aéreas para criação e indicação de pacotes de viagem ou voos promocionais, ou pelos departamentos de turismo das cidades, para criação de programas de incentivo ao turismo e aumento da disponibilidade de atrações turísticas. Empresas de transporte particular têm também a chance de oferecer traslados para eventos nas cidades de interesse, bem como empresas de viação terrestre podem reduzir os valores das passagens ou criar rotas mais interessantes para os viajantes.

É importante ressaltar que o escopo deste trabalho envolveu a análise dos dados dos perfis de usuários que visitaram a cidade de Curitiba. Entretanto, este tipo de análise não se restringe apenas a Curitiba. As condições analisadas tanto nesta seção quanto ao longo deste estudo são aplicáveis a quaisquer outras cidades. Além disso, os resultados desta seção, assim como das seguintes, evidenciam a importância do papel da computação quando se trata de análise de grandes volumes de dados. Como dito no início desta seção,

Figura 15: Número de ocorrências das capitais dos estados

foram considerados 42.381 perfis de usuários. Coletar, limpar e analisar as informações de todos estes perfis manualmente seria uma tarefa bastante complicada, senão impossível, e poderia levar meses. Com o auxílio da tecnologia, entretanto, esse processo pôde ser concluído em poucos dias.

As seções a seguir mostram os dados analisados e as possíveis aplicações das informações obtidas a partir das *tags* e *badges*.

5.1.2 ANÁLISE DAS *TAGS* DE ESTILO DE VIAGEM

Como já explicado na seção 3, as *tags* são rótulos que ajudam a definir os estilos preferidos de viagem dos usuários do TripAdvisor.

Dos quase 43 mil perfis analisados, foi observado que a maioria dos viajantes, 12.655 (10.4%), gosta da cultura local. Este número, apesar de ser o mais acentuado, não se distancia muito da quantidade de pessoas que definiram um estilo de viagem *Gourmet* (10.458 - 8.6%), nem das que afirmaram adorar praia (10.363 - 8.5%) e viajar com a família (10.205 - 8.4%). Além disso, 8.717 (7.2%) disseram ser amantes da natureza e, em número

muito similar, 8.712 (7.1%) buscam paz e tranquilidade. Os que economizam na viagem somam 7.548 (6.2%) ao mesmo tempo que os que adoram fazer compras são 7.488 (6.1%). Foram totalizados 7.408 (6.1%) fãs de história e 7.348 (6%) usuários que adoram a vida urbana. Os fãs de arte e arquitetura somam 5.337 (4.4%), os que buscam emoção, 4.970 (4.1%), os viajantes formadores de opinião, 4.749 (3.9%) e os que adoram vida noturna, 4.021 (3.3%). São 3.669 (3%) os que se identificam como ecoturistas, 3.503 (2.9%) os que adoram luxo e 3.102 (2.5%) os viajantes mochileiros. Aqueles com mais de 60 anos, 958, e vegetarianos, 639, representam 0.8% e 0.5% do total analisado, respectivamente. Essas informações podem ser melhor visualizadas no gráfico 16 e nuvem de palavras a seguir (figura 17).

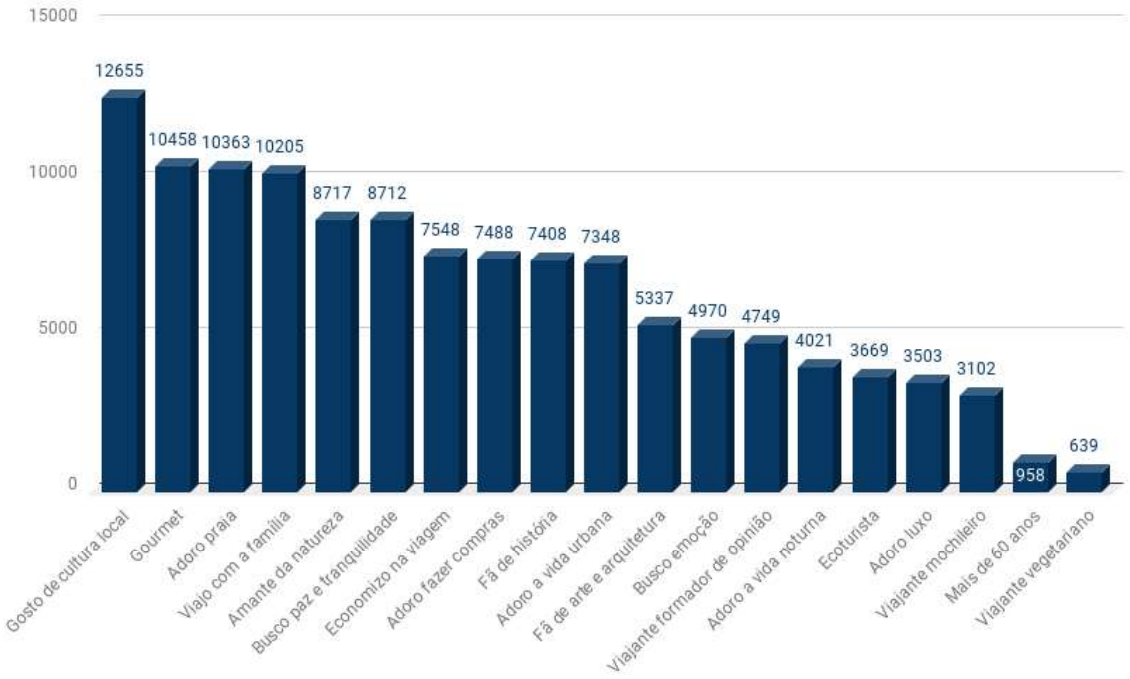

Figura 16: Gráfico da ocorrência de *tags*

Diferentemente das informações sobre as cidades, as estatísticas obtidas para as *tags* de estilo de viagem não retratam para onde os turistas gostam de viajar, mas sim o tipo de viagem que preferem. Sabendo que a maioria deles gosta de cultura local e apontaram um estilo *Gourmet*, por exemplo, as prefeituras e casas de eventos têm a possibilidade de trazer ou divulgar mais eventos culturais nas cidades. Além disso, restaurantes podem criar promoções próprias ou se vincular a festivais gastronômicos, a exemplo dos que já acontecem em Curitiba, como o Festival Bom Gourmet (POVO, 2018). De modo geral, as informações fornecidas pelas *tags*, por retratarem os estilos preferidos

Figura 17: Nuvem de palavras da ocorrência de *tags*

de viagem dos turistas, ou seja, características do perfil dos usuários, apresentam grande potencial para serem utilizadas como *features* em sistemas de recomendação.

5.1.3 ANÁLISE DE *BADGES*

Quanto aos *badges*, o gráfico da figura 18 mostra a frequência com que foram concedidos aos usuários e a nuvem de palavras (figura 19) ajuda a perceber mais claramente quais são os mais frequentes.

O selo mais relevante é o de Colaborador Importante, que aparece 34.633 vezes. Logo atrás vem o selo Passaporte, com 33.346 aparições. Em seguida vem os selos de especialidade: Expert em restaurantes (27.194), Expert em atrações (26.231), Expert em hotéis (25.587) e o selo Explorador (22.176), dado quando o colaborador é um dos primeiros a fazer a avaliação de um lugar em um determinado idioma. Os demais selos não apresentam grandes destaques, e a lista é composta por: Expert em pousadas (5.910), Expert em hotéis luxuosos (4.485), Colaborador do Travellers' Choice 2015 para hotéis (4.233), Colaborador do Travellers' Choice 2016 para hotéis (3.037), Expert em resorts (2.466), Colaborador do Travellers' Choice 2015 para atrações (1.582) e Expert em hotéis-boutique (1.029).

Nota-se que, dentre as especialidades, destacam-se Expert em restaurantes, atrações e hotéis, especificamente nesta ordem. O selo Colaborador Importante possui níveis e é

concedido de acordo com a quantidade de votos úteis dados pelo usuário, conforme mostrado no tópico Votos Úteis da seção 3. Nesta análise os níveis foram abrigados e são apresentados apenas como Colaborador Importante, pois cada nível isolado representaria resultados pouco significativos. De modo geral, é possível perceber que há um interesse dos colaboradores em divulgar, principalmente, informações sobre os locais que visitam, os restaurantes que experimentam e os hotéis em que se hospedam.

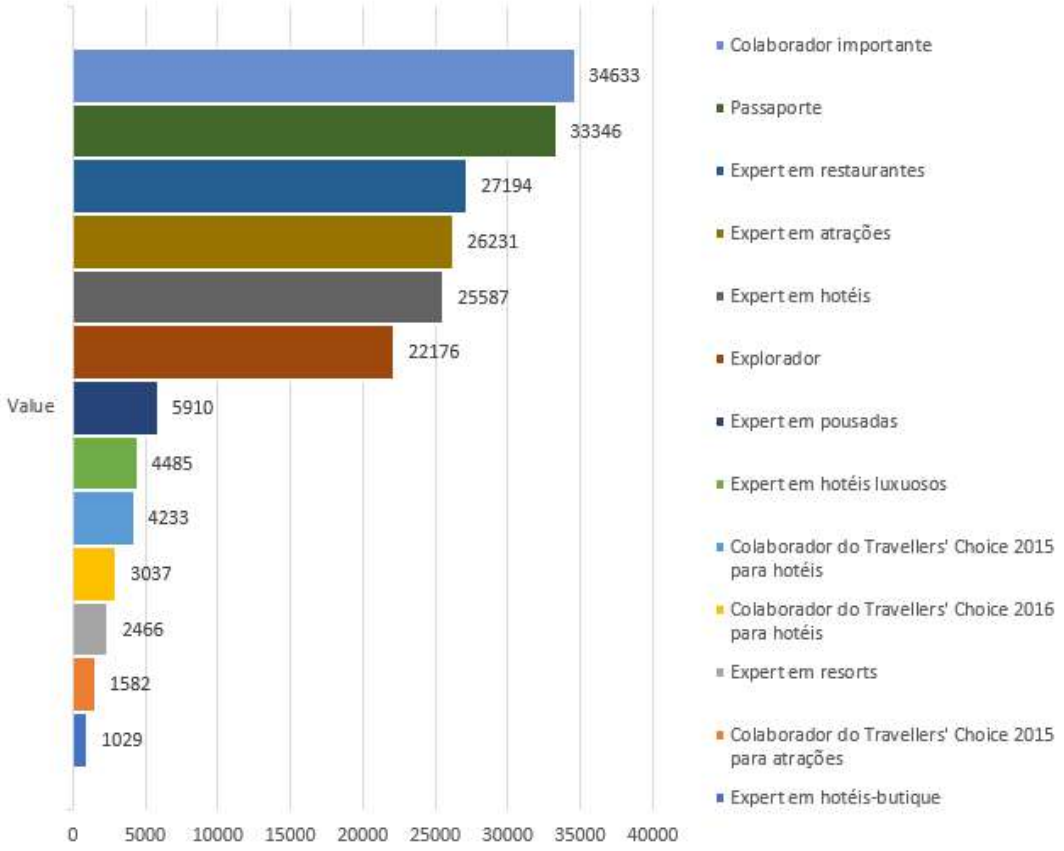

Figura 18: Gráfico da quantidade de *badges*

5.2 PERFIS DOS USUÁRIOS - APLICAÇÃO DO ALGORITMO *APRIORI*

Assim como os resultados mostrados até aqui, os que são apresentados nesta seção também foram coletados a partir das cidades avaliadas, das *tags* de estilo de viagem e dos *badges*, porém aplicados ao algoritmo *apriori*. Como visto na seção 4.5, o *apriori*, quando aplicado a um *dataset*, retorna um conjunto de regras de associação entre os elementos (cidades, *tags* e *badges*, no caso deste trabalho). Para cada um dos atributos foram definidos os valores de confiança e suporte, utilizados para encontrar regras de associação válidas e significativas, conforme explicado na seção 4.5. Os valores desses argumentos foram

Figura 19: Nuvem de palavras da ocorrência de *badges*

definidos testando-se diferentes combinações de valores. Os que apresentaram os melhores resultados foram então escolhidos. Sendo assim, tais resultados serão apresentados a seguir.

5.2.1 APLICAÇÃO DO *APRIORI* SOBRE O CONJUNTO DE CIDADES

A tabela 1 mostra as regras de associação entre as cidades presentes na base de dados coletada do TripAdvisor. Os valores de confiança e suporte utilizados para obtê-las foram de 0.9 e 0.1, respectivamente. As regras representam uma relação A $\Rightarrow$ B que, neste contexto, refere-se a "quem visitou a cidade A visitou também a cidade B". De acordo com esta tabela, podemos notar que todas as 17 cidades presentes na posição de A tem uma relação de associação com a cidade de Curitiba, ocupante da posição B. Isso porque, como já dito anteriormente, a base de dados utilizada foi a de ações realizadas a partir da capital paranaense. Dentre estas cidades, estão 13 capitais, sendo 9 nacionais e 4 internacionais. As demais cidades, Balneário Camboriú, Orlando, Foz do Iguaçú e Gramado, apesar de não serem capitais, são muito famosas por seus atrativos turísticos.

Sobre estas últimas, é fácil pensar em hipóteses sobre a relação delas com Curitiba. Balneário Camboriú é uma das praias de Santa Catarina mais frequentadas por curitibanos, principalmente em épocas de alta temporada, estando localizada a menos de 3 horas da capital. Além disso, turistas que vêm de cidades mais distantes para Balneário

acabam muitas vezes por fazer uma parada em Curitiba. Orlando foi considerado, entre os anos de 2015 e 2017, um dos destinos americanos mais populares entre os brasileiros (DISNEY, 2015) (DISNEY, 2016) (PAUTA, 2017). Isso se dá principalmente pelas atrações de parques temáticos que a cidade oferece, mas também provavelmente pelo valor dos bilhetes aéreos, que tornam o destino uma opção financeiramente viável. Além disso, Miami, que é cidade próxima e comummente visitada junto com Orlando, teve, até 2016, voos diretos para Curitiba, o que possivelmente influenciou a frequência e quantidade de passageiros que visitavam Orlando e passavam por Curitiba. Essa hipótese pode ser reforçada pelas declarações da American Airlines de que "o desempenho das vendas para o voo inaugural e as primeiras semanas de operação surpreenderam os executivos"(FRANCO, 2014) e que "o desempenho foi considerado o melhor já realizado para o lançamento de um novo trecho da empresa no país"(FRANCO, 2014). Em Foz do Iguaçú encontram-se as Cataratas do Iguaçú, destino turístico mundialmente famoso e que contempla a lista das sete maravilhas da natureza segundo a *New Open World Corporation* (WONDERS, 2011). Assim como Curitiba, está localizada no estado do Paraná, tendo o aeroporto da capital como um dos mais próximos. Já Gramado é um destino muito popular durante o inverno, quando é famoso por seus rodízios de *foundue* e durante os meses de outubro a janeiro, época do Natal Luz, quando a cidade oferece diversas atrações natalinas aos turistas (GRAMADOTUR, 2018).

As demais regras podem ser explicadas considerando fatores como distância, rotas de voos e até semelhanças nos estilos das cidades. Seja como for, as regras geradas ajudam a compreender o comportamento dos usuários e, com base nisso, é possível tomar medidas direcionadas de marketing e vendas com sugestões de rotas de viagem, por exemplo.

5.2.2 APLICAÇÃO DO *APRIORI* SOBRE O CONJUNTO DE *TAGS*

A tabela 2 mostra as 11 regras de associação encontradas. Neste caso a relação A $\Rightarrow$ B é referente a "quem escolheu a *tag* A escolheu também a *tag* B". É possível notar que, dentre as 19 opções de *tags* disponíveis, apenas 8 tem associações relevantes entre si. Para esta análise foram utilizados confiança e suporte com valores de 0.6 e 0.1, respectivamente.

Algumas das regras são bastante intuitivas, por exemplo a primeira. De acordo com o dicionário Michaelis, *Gourmet* é uma palavra originária do francês que, significa "Pessoa que é grande conhecedora e apreciadora de boa comida e bons vinhos"(MICHAELIS, 2018). Luxo, também de acordo com o dicionário, tem como duas de suas definições "coisa

Tabela 1: Regras de associação entre o conjunto de cidades

regra	lhs	rhs	suporte	confiança	lift
1	Natal	⇒ Curitiba	0.1024027	0.9905200	1.0270624
2	Balneário Camboriú	⇒ Curitiba	0.1147872	0.9643214	0.9998973
3	Nova York	⇒ Curitiba	0.1070357	0.9939327	1.0306011
4	Orlando	⇒ Curitiba	0.1183214	0.9955022	1.0322285
5	Fortaleza	⇒ Curitiba	0.1162128	0.9673671	1.0030554
6	Salvador	⇒ Curitiba	0.1096790	0.9435360	0.9783451
7	Belo Horizonte	⇒ Curitiba	0.1182917	0.9402738	0.9749626
8	Brasília	⇒ Curitiba	0.1290428	0.9935971	1.0302530
9	Paris	⇒ Curitiba	0.1096196	0.9456828	0.9805711
10	Santiago	⇒ Curitiba	0.1271718	0.9579418	0.9932824
11	Porto Alegre	⇒ Curitiba	0.1362894	0.9774228	1.0134821
12	Foz do Iguaçú	⇒ Curitiba	0.1824716	0.9956247	1.0323555
13	Gramado	⇒ Curitiba	0.1835704	0.9596336	0.9950366
14	Florianópolis	⇒ Curitiba	0.1977369	0.9952167	1.0319325
15	Buenos Aires	⇒ Curitiba	0.1778979	0.9567162	0.9920115
16	Rio de Janeiro	⇒ Curitiba	0.3001396	0.9909786	1.0275380
17	São Paulo	⇒ Curitiba	0.4040569	0.9580311	0.9933750
18	Florianópolis,São Paulo	⇒ Curitiba	0.1066199	0.9950111	1.0317192
19	Buenos Aires,São Paulo	⇒ Curitiba	0.1049865	0.9608589	0.9963071
20	Rio de Janeiro,São Paulo	⇒ Curitiba	0.1745716	0.9897289	1.0262422

excelente, de qualidade superior; perfeição, primor"e "qualquer coisa dotada de excelência no mais alto grau"(MICHAELIS, 2018). Tais definições ajudam a endossar a regra {Adoro luxo} ⇒ {Gourmet}, que mostra um perfil de viajantes que busca qualidade e, portanto, também gosta de comer bem. Outras regras, como {Adoro a vida noturna} ⇒ {Adoro praia}, não seriam tão simples de prever sem a ajuda de uma ferramenta de análise de dados.

Os resultados da aplicação do *apriori* sobre o conjunto de *tags* evidenciam o potencial deste atributo como ferramenta para tomada de decisões de negócios. As informações obtidas a partir desta análise podem servir como diretriz para a elaboração de pacotes turísticos mais direcionados por parte de agências de viagens, por exemplo.

5.2.3 APLICAÇÃO DO *APRIORI* SOBRE O CONJUNTO DE *BADGES*

Para o conjunto de *badges* utilizou-se para os parâmetros confiança e suporte os valores 0.5 e 0.8, respectivamente. Para o contexto dos *badges* a relação A ⇒ B é referente a "quem recebeu o selo A recebeu também o selo B". Ao todo foram geradas 17 regras, conforme mostrado na tabela 3. Foi necessário diminuir o tamanho da tabela

Tabela 2: Regras de associação entre *tags*

regra	lhs	rhs	suporte	confiança	lift
1	Adoro luxo	⇒ Gourmet	0.1087	0.7159	1.579
2	Ecoturista	⇒ Adoro praia	0.1004	0.6317	1.406
3	Ecoturista	⇒ Gosto de cultura local	0.1111	0.6985	1.387
4	Adoro a vida noturna	⇒ Adoro praia	0.1122	0.6734	1.498
5	Adoro a vida noturna	⇒ Gourmet	0.1044	0.6271	1.383
6	Adoro a vida noturna	⇒ Gosto de cultura local	0.1102	0.6617	1.314
7	Adoro a vida urbana	⇒ Gosto de cultura local	0.1439	0.6212	1.234
8	Busco emoção	⇒ Adoro praia	0.1412	0.6553	1.458
9	Busco emoção	⇒ Gosto de cultura local	0.1415	0.6567	1.304
10	Adoro praia,Gourmet	⇒ Gosto de cultura local	0.1523	0.6474	1.286
11	Adoro praia,Gosto de cultura local	⇒ Gourmet	0.1523	0.6144	1.355

de acordo com o tamanho da página e, por esse motivo, os títulos dos *badges* “Expert em hotéis”, “Expert em atrações”e “Expert em restaurantes”foram reduzidos para “Hotéis”, “Atrações”e “Restaurantes”, respectivamente.

Comparando os atributos do TripAdvisor entre si, estes nos permitem perceber que: as cidades descrevem para onde os turistas mais gostam de viajar; as *tags* definem o estilo de viagem que preferem; já os *badges* mostram o que estes usuários mais se dispõem a avaliar. No caso deste último, nota-se, por exemplo, que dentre as opções de hospedagem (pousadas, hotéis boutique, resorts, hotéis luxuosos), a que se destaca são apenas hotéis. Também é possível perceber que há um interesse dos viajantes em serem os pioneiros na avaliação de algum lugar em determinado idioma, dada a relevância do selo Explorador.

Por ser um atributo relativo a revisões e exposição da opinião dos turistas para leitores do mundo inteiro, este tipo de funcionalidade pode atuar não só como incentivo para os usuários continuarem contribuindo e evoluindo em seus perfis, mas também para empresas prestadoras de serviços melhorarem seu atendimento aos clientes, com o intuito de evitar revisões negativas e ampliar a quantidade de positivas. Dessa forma é possível observar que, assim como cidades e *tags*, os *badges* também tem um papel significativo na tarefa de identificar os perfis de usuários do TripAdvisor.

Tabela 3: Regras de associação entre *badges*

regra	lhs	rhs	suporte	confiança	lift
1	Explorador	⇒ Hotéis	0.5341443	0.8553391	1.187085
2	Explorador	⇒ Atrações	0.5358340	0.8580447	1.161604
3	Explorador	⇒ Restaurantes	0.5695418	0.9120220	1.190954
4	Hotéis	⇒ Atrações	0.6018980	0.8353461	1.130875
5	Atrações	⇒ Hotéis	0.6018980	0.8148374	1.130875
6	Hotéis	⇒ Restaurantes	0.6264256	0.8693868	1.135280
7	Restaurantes	⇒ Hotéis	0.6264256	0.8180113	1.135280
8	Atrações	⇒ Restaurantes	0.6461378	0.8747284	1.142255
9	Restaurantes	⇒ Atrações	0.6461378	0.8437523	1.142255
10	Hotéis,Explorador	⇒ Restaurantes	0.5024077	0.9405841	1.228252
11	Restaurantes,Explorador	⇒ Hotéis	0.5024077	0.8821261	1.224262
12	Hotéis,Restaurantes	⇒ Explorador	0.5024077	0.8020229	1.284300
13	Atrações,Explorador	⇒ Restaurantes	0.5092507	0.9503889	1.241055
14	Restaurantes,Explorador	⇒ Atrações	0.5092507	0.8941409	1.210470
15	Atrações,Hotéis	⇒ Restaurantes	0.5617696	0.9333302	1.218779
16	Hotéis,Restaurantes	⇒ Atrações	0.5617696	0.8967858	1.214051
17	Atrações,Restaurantes	⇒ Hotéis	0.5617696	0.8694269	1.206637

6 CONCLUSÃO E TRABALHOS FUTUROS

Como foi mostrado ao longo deste trabalho, por possibilitar comentários e notas em pontos turísticos, hotéis, pousadas, restaurantes e atrações, o TripAdvisor reúne diversos tipos de informação sobre muitas cidades do mundo. Mais do que isso, reúne dados de seus usuários. Neste estudo foram explorados os atributos particulares desta plataforma, *tags* e *badges*, bem como as cidades avaliadas pelos colaboradores de Curitiba. A pesquisa, de caráter exploratório, buscou compreender as características de perfil destes usuários do TripAdvisor a partir da análise de tais atributos. Para tanto, utilizou-se o algoritmo de mineração de dados *apriori*, a fim de encontrar regras de associação entre os registros de cidades visitadas, entre os selos colecionados e entre *tags* escolhidas.

A realização deste trabalho permitiu que fosse reafirmado o potencial da computação na análise de grandes volumes de dados. A base coletada e utilizada foi de 42.381 usuários, e seria tarefa bastante complexa coletar, armazenar e analisar tantas informações sem o apoio de ferramentas computacionais. Além disso, neste estudo optou-se por utilizar apenas dados de Curitiba, mas sabe-se que estender estas análises para quaisquer outras cidades também seria possível graças à tecnologia empregada.

Quanto aos resultados obtidos, estes reforçam a hipótese inicial de que os atributos do TripAdvisor seriam fontes valiosas de informações que nos ajudariam a compreender melhor o perfil dos usuários. Por meio destes resultados foram descobertas as cidades para onde as pessoas que visitaram Curitiba mais viajam, quais são seus estilos de viagem preferidos e os *badges* mais concedidos a elas. Tomando como base estas análises, são diversas as possibilidades de aplicação e de desenvolvimento de outros estudos que aprofundem tópicos mais específicos dentro das mais diversas linhas de pesquisa.

Como já sugerido, uma das possibilidades de trabalhos futuros é a expansão do escopo deste estudo para mais cidades do planeta. Agrupar os usuários em classes, ou "tribos", também pode trazer *insights* diferenciados e contribuir para um melhor entendimento a respeito de seus perfis. Os resultados obtidos têm potencial para serem

utilizados como base para o desenvolvimento de estudos que se apliquem em diferentes níveis de serviços de turismo, como a elaboração de sistemas de recomendação de viagens, seja de pacotes ou de roteiros turísticos, ou o apoio em tomadas de decisões durante o planejamento de viagens. As aplicações deste conhecimento, entretanto, não se limitam somente à área da tecnologia. Entender o perfil dos usuários pode contribuir também para o estudo e a realização de ações de marketing direcionadas, por exemplo.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: 20TH INT'L CONFERENCE ON VERY LARGE DATABASES. [S.l.], 1994.

ALBUQUERQUE, E. M. **O que faremos com os 40 trilhões de gigabytes de dados disponíveis em 2020?** Setembro 2017. Disponível em: <https://br.okfn.org/2017/09/29/o-que-faremos-com-os-40-trilhoes-de-gigabytes-de-dados-disponiveis-em-2020/>.

AMARAL, F.; TIAGO, T.; TIAGO, F. User-generated content: tourists' profiles on tripadvisor. **International Journal on Strategic Innovative Marketing**, 2014.

AYEH, J. K.; AU, N.; LAW, R. Do we believe in tripadvisor?: examining credibility perceptions and online travelers' attitude toward using user-generated content. **Journal of travel research : a quarterly publication of the Travel and Tourism Research Association**, v. 52, n. 4, 2013.

BALUJA, S. et al. Video suggestion and discovery for youtube: Taking random walks through the view graph. In: WORLD WIDE WEB CONFERENCE (WWW). [S.l.], 2008. p. 895–904.

BARAGLIA, R. et al. Learnext: learning to predict tourists movements. In: CIKM. [S.l.], 2013. p. 622–625.

BATET, M. et al. Turist@: Agent-based personalised recommendation of tourist activities. **Expert Systems with Applications: An International Journal**, v. 39, n. 8, 2012.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Coleta e análise de grandes bases de dados de redes sociais online. In: UFMG - UNIVERSIDADE FEDERAL DE MINAS GERAIS. [S.l.], 2012.

BIDART, R. Uso de informações específicas de domínio em recomendações para turismo. **Dissertação (mestrado) — Universidade Federal de Minas Gerais**, 2015.

BRILHANTE, I. R. et al. Where shall we go today?: planning touristic tours with tripbuilder. In: ACM. [S.l.], 2013. p. 757–762.

BURKLE, M.; MARLOW, C.; LENTO, T. Feed me: Motivating newcomer contribution in social network sites. 2009.

CASTILLO, C. Effective web crawling. In: ACM. **ACM SIGIR Forum**. [S.l.], 2005. v. 39, n. 1, p. 55–56.

CHENG, Z. et al. Exploring millions of footprints in location sharing services. In: ICWSM. [S.l.], 2011. p. 81–88.

CHEONG, F.-C. **Internet agents: spiders, wanderers, brokers, and bots**. [S.l.]: New Riders Publishing, 1996.

CHOUDHURY, M. D.; FELDMAN, M.; AMER-YAHIA, S. Automatic construction of travel itineraries using social breadcrumbs. In: HT. [S.l.], 2010.

CHUN, H. et al. Comparison of online social relations in volume vs interaction: a case study of cyworld. In: ACM SIGCOMM INTERNET MEASUREMENT CONFERENCE (IMC). [S.l.], 2008. p. 57–70.

CIO. **Tome nota: 2,5 quintilhões de bytes são criados todos os dias**. Outubro 2015. Disponível em: <http://cio.com.br/noticias/2015/10/27/tome-nota-2-5-quintilhoes-de-bytes-sao-criados-todos-os-dias/>.

CROCKFORD, D. The application/json media type for javascript object notation (json). 2006.

DISNEY, C. na. **Gramado e Orlando são os destinos mais procurados pelos brasileiros, segundo TripAdvisor**. [S.l.]: Casa na Disney, May 2015.

DISNEY, C. na. **Orlando é um dos destinos preferidos de quem viajou ao exterior em 2016**. [S.l.]: Casa na Disney, 2016.

DUARTE, F. et al. Traffic characteristics and communication patterns in blogosphere. In: WEBLOGS AND SOCIAL MEDIA (ICWSM). [S.l.], 2007.

FARRAHI, K.; GATICA-PEREZ, D. Discovering routines from large-scale human locations using probabilistic topic models. In: ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY. [S.l.], 2011.

FERREIRA, A. P. G.; SILVA, T. H.; LOUREIRO, A. A. F. Beyond sights: Large scale study of tourists' behavior using foursquare data. In: IEEE ICDM - MOBILITY ANALYTICS FROM SPATIAL AND SOCIAL DATA. [S.l.], 2015.

FRANCO, A. P. **Voo direto de Curitiba para Miami tem novo horário**. [S.l.]: Gazeta do Povo, March 2014.

GRAMADOTUR. **Natal Luz Gramado - Programação Oficial**. 2018. Disponível em: <http://www.natalluzdegramado.com.br/programacao/>.

GRETZEL, U.; YOO, K. H.; PURIFOY, M. Online travel review study: role & impact of online travel reviews. **Laboratory for Intelligent Systems in Tourism**, 2007.

GUROVITZ, H. **O que cerveja tem a ver com fraldas?** February 2011. Disponível em: <https://exame.abril.com.br/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/>.

HAN, J.; KRAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**. [S.l.]: Elsevier Inc., 2012.

HECHT, B. et al. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In: SIGCHI '11. [S.l.], 2011.

HUANG, Y.; BIAN, L. A bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet. **Expert Systems with Applications**, v. 36, n. 1, 2009.

KIM, C. E-tourism: an innovative approach for the small and medium-sized tourism enterprises (smtes) in korea. **Organisation for Economic Co-operation and development (OECD)**, 2004.

LINDQVIST, J. et al. I'm the mayor of my house: Examining why people use foursquare - a social-driven location sharing application. In: SIGCHI '11. [S.l.], 2011.

LONG, X.; JIN, L.; JOSHI, J. Towards understanding traveler behavior in location-based social networks. In: GLOBECOM. [S.l.], 2013.

LU, W.; STEPCHENKOVA, S. User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. **Journal of Hospitality Marketing & Management**, 2014.

MACHADO, C. C. et al. Um web crawler para projeções e análise de vulnerabilidades de segurança e consistência estrutural de páginas web. **Revista de Empreendedorismo, Inovação e Tecnologia**, v. 2, n. 2, p. 3–12, 2016.

MAJID, A. et al. Gothere: Travel suggestions using geotagged photos. In: ACM. [S.l.], 2012. p. 577–578.

MARTINS, T. **Estudo mundial levanta os dados da Internet no Brasil e no mundo, descubra as principais redes sociais e comportamento de compras online dos usuários**. Abril 2017. Disponível em: <http://marketingsemgravata.com.br/site/2017/04/17/dados-da-internet-2017-brasil-redes-sociais/>.

MICHAELIS. **Michaelis Online**. Editora Melhoramentos, 2018. Disponível em: <http://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/gourmet/>.

MISLOVE, A. et al. Measurement and analysis of online social networks. In: . [S.l.: s.n.], 2007.

MUNAR, A. M.; GYIMOTHY, S.; CAI, L. **Tourism Social Media: Transformations in Identity, Community and Culture**. [S.l.]: Emerald, 2013.

O'CONNOR, P. User-generated content and travel: A case study on tripadvisor. com. **Information and communication technologies in tourism 2008**, Springer, p. 47–58, 2008.

PAUTA, V. em. **Conheça os 10 destinos internacionais mais procurados por brasileiros, em 2017**. [S.l.]: Viagem em Pauta, December 2017.

PELECHRINIS, K.; KRISHNAMURTHY, P. Location-based social network users through a lense: Examining temporal user patterns. In: AAAI SNSC 2012 (FALL SYMPOSIUM). [S.l.], 2012.

PESSOA, B. C. et al. Banco de dados mongodb vs banco de dados sql server 2008. **RE3C-Revista Eletrônica Científica de Ciência da Computação**, v. 7, n. 1, 2012.

PIANESE, F.; KAWSAR, F.; ISHIZUKA, H. Discovering and predicting user routines by differential analysis of social network traces. In: WOWMOM. [S.l.], 2013. p. 1–9.

POPESCU, A.; GREFENSTETTE, G. Mining social media to create personalized recommendations for tourist visits. In: ACM. [S.l.], 2011.

POVO, G. do. **Festival Bom Gourmet Gazeta do Povo**. 2018. Disponível em: <http://www.gazetadopovo.com.br/bomgourmet/festivais/festival-bom-gourmet>.

ROY, S. B. et al. Interactive itinerary planning. In: ICDE. [S.l.], 2011. p. 15–26.

RUSSO, I. **Agências oferecem viagens sob medida para diferentes tipos de turistas**. Folha de S. Paulo, 2016. Disponível em: $<http://www1.folha.uol.com.br/turismo/2016/11/1832658-agencias-oferecem-viagens-sob-medida-para-diferentes-tipos-de-turistas.shtml?cmpid=compfb>$.

SANTOS, F. A. et al. Towards a sustainable people-centric sensing. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS. [S.l.], 2017.

SHI, Y. et al. Personalized landmark recommendation based on geotags from photo sharing sites. In: ICWSM. [S.l.], 2011. p. 622–625.

TRIPADVISOR. **Fact sheet - TripAdvisor**. 2015. Disponível em: $<http://www.tripadvisor.com/PressCenter-c4-FactSheet.html>$.

TRIPADVISOR. **Perguntas Frequentes do TripColaboradores**. 2016. Disponível em: $<https://www.tripadvisor.com.br/TripCollectiveFAQ>$.

TRIPADVISOR. **Selos TripColaboradores: o que são e como faço para recebê-los?** 2016. Disponível em: $<https://www.tripadvisor.com.br/TripCollectiveBadges>$.

TRIPADVISOR. **TripColaboradores - TripAdvisor**. 2016. Disponível em: $<https://www.tripadvisor.com.br/TripCollective>$.

TRIPADVISOR, L. **Content API**. 2018. Disponível em: <https://developer-tripadvisor.com/content-api/>.

UWNTO. **United nations world tourism organization**. 2014. Disponível em: <http://media.unwto.org/press-release/2015-04-15/ exports-international-tourism-rise-us-15-trillion-2014>.

VASCONCELOS, L. M. R. d.; CARVALHO, C. L. d. Aplicação de regras de associação para mineração de dados na web. In: INSTITUTO DE INFORMáTICA - UNIVERSIDADE FEDERAL DE GOIáS. [S.l.], 2004.

WALTER, R.; DUARTE, D. Implementaç ao de operadores olap utilizando o modelo de programaç ao map reduce no mongodb. 2016.

WASKO, M. M.; FARAJ, S. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. **MIS Quarterly**, v. 29, n. 1, p. 35, 2005.

WASSON, M.; TEJADA, Z. **Non-relational data and NoSQL**. Microsoft Azure, November 2017. Disponível em: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/non-relational-data>.

WONDERS, N. . **New 7 Wonders of nature**. Global Platform Ltd, 2011. Disponível em: <https://nature.new7wonders.com/>.

WRITEWORDS. **Word Frequency Counter**. 2018.

YERVA, S. R. et al. Tripeneer: User-based travel plan recommendation application. In: ICWSM. [S.l.], 2013.

YOON, H. et al. Smart itinerary recommendation based on user-generated gps trajectories. In: INTEL. AND COMPUTING. [S.l.], 2010. p. 19–34.

ZHENG, Y. Para onde devo viajar: Recomendação de cidades baseada em comunidades de usuários. In: BRASNAM. [S.l.], 2014.