Mineração de conteúdos em **mídias sociais**: Uma construção metodológica[1]

GUILHERME MENDES THOMAZ * [guimendesthomaz@gmail.com]
EDUARDO MICHELOTTI BETTONI ** [eduardo@odois.org]
ALEXANDRE AUGUSTO BIZ *** [biz@ufpr.br]
CLARA KAMILA BATISTA SANTOS **** [kllara10@gmail.com]

Resumo | No presente artigo objetivou-se apresentar as metodologias de mineração de conteúdos em mídias sociais e seus os desafios, além de analisar modelos de descoberta de conhecimento (*Knowledge Discovery in Databases* – KDD). As mídias sociais provocaram mudanças significativas na atividade turística, principalmente em relação aos conteúdos gerados por usuários (*User Generated Conten* – UGC). A expansão exponencial de dados online fez com que as mídias sociais se tornassem uma potencial fonte de informações e conhecimentos relevantes para a tomada de decisão organizacional. Nesse cenário, a mineração de conteúdo é considerada uma atividade desafiadora e exige o desenvolvimento de novas técnicas e ferramentas capazes de coletar e transformar, de forma rápida e automatizada, um grande volume de dados em conhecimento. Este estudo exploratório de base documental e bibliográfica foi alicerçado em artigos científicos das áreas de Ciência e Gestão da Informação, Sistemas de Informação, Tecnologia e Internet, Ciência da Computação, Recuperação de Informação e Mineração de Dados. O resultado foi a proposição de uma metodologia de mineração de conteúdos em mídias sociais para auxiliar na gestão de destinos turísticos.

Palavras-chave | Mineração de conteúdos em mídias sociais, Mídias sociais, Metodologia de pesquisa, Pesquisa em turismo.

Abstract | The present paper aims to present social media content mining methodologies, its challenges and analyze Knowledge Discovery in Databases (KDD) models. Social media has brought significant changes in tourism, particularly the User Generated Content (UGC). With the exponential growth of online data, social media has become a potential source of relevant information and knowledge to organizational decision making. In this scenario, content mining is considered a challenging activity that requires the development of new techniques and tools to collect and transform in a fast and automated manner, a large volume of data into knowledge. This exploratory study was grounded in papers and researches of Science and Information Management, Information Systems, Technology and Internet, Computer Science, Information Retrieval and Data Mining areas. The result was the development of a social media content mining methodology to assist in the management of tourist destinations.

Keywords | Social media mining, Social media, Research methodology, Research in tourism.

[1] Pesquisa financiada por MCTI/CNPq/MEC/CAPES n. 18/2012 e CNPq/COENG – Copa do Mundo 2014.

* **Mestrando em Turismo (PPGTUR)** na Universidade Federal do Paraná (UFPR). **Bolsista** do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e Pesquisador do Laboratório de Turismo, Tecnologia, Informação, Comunicação e Conhecimento (TURITEC) Brasil.

** **Mestre em Ciência, Gestão e Tecnologia da Informação** pela Universidade Federal do Paraná (UFPR). **Pesquisador** dos Observatórios Sesi/Senai/IEL e do Laboratório de Turismo, Tecnologia, Informação, Comunicação e Conhecimento (TURITEC), Brasil.

*** **Pós-Doutor Empresarial PDI-CNP** realizado no Instituto Stela (Florianópolis/SC) na área Gestão do Conhecimento, e **Pós-Doutor** em Gestão do Conhecimento pelo Instituto Stela (Florianópolis/SC) pela Universidade Federal de Santa Catarina (UFSC). **Professor** do Departamento de Turismo da Universidade Federal do Paraná (UFPR), e **Pesquisador** do Laboratório de Turismo, Tecnologia, Informação, Comunicação e Conhecimento (TURITEC), Brasil.

**** **Bacharel em Turismo** pela Universidade Federal do Paraná (UFPR) e **Pesquisadora** do Laboratório de Turismo, Tecnologia, Informação, Comunicação e Conhecimento (TURITEC), Brasil.

1. Introdução

A popularização da chamada *web* interativa ou 2.0, principalmente no formato de mídias sociais, ocasionou transformações significativas na atividade turística, principalmente em relação a produção e compartilhamento de conteúdo *online* por viajantes (Xiang & Gretzel, 2010).

As informações são publicadas e compartilhadas pelos usuários nas mídias sociais em forma de texto, imagens, vídeos e áudios; envolvendo fatos, opiniões, avaliações, impressões, sentimentos, rumores e experiências que são criadas, iniciadas, compartilhadas e utilizadas entre usuários em relação a produtos, marcas, serviços e problemas (Torres, 2009; Blackshaw & Nazzaro, 2006).

A ampla adoção das mídias sociais por usuários tem gerado uma expansão exponencial de conteúdo onde existem, ainda que não em totalidade, informações relevantes para determinados atores sociais – como é o caso das organizações. Em vários formatos, a informação pode auxiliar no processo decisório, especialmente no planejamento e na gestão, desde que seu significado ou valor possa ser identificado em meio a vastidão de conteúdo disponível. Sequer os sistemas convencionais ou banco de dados de parte significativa das empresas tem capacidade para lidar com tal volume, configurando-se como um desafio no desenvolvimento de técnicas e ferramentas adequadas. Ao efetivo controle e trato dessa problemática é utilizado o termo mineração de conteúdo em mídias sociais, ou *Social Media Mining* (SMM) (Dey, Haque, Khurdiya & Shroff, 2011; Governatori & Iannella, 2011).

Grande parte das organizações não está familiarizada ou não tem conhecimento suficiente desse processo. O problema sequer se concentra na obtenção dos dados, mas sim, no seu tratamento de modo que sirvam à objetivos operacionais e estratégicos (Dai, Kakkonen & Sutinen, 2011).

Sendo assim, utilizando-se de uma pesquisa exploratória tendo a revisão bibliográfica como método de investigação; no presente artigo objetivou-se apresentar as metodologias de SMM e seus desafios, além de analisar modelos de descoberta de conhecimento (*Knowledge Discovery in Databases – KDD*); sendo estas etapas para o desenvolvimento de uma metodologia de mineração de conteúdos em mídias sociais que auxilie na gestão de destinos turísticos.

2. Mineração de conteúdo em mídias sociais e suas oportunidades na atividade turística

O termo SMM é uma variação de *web mining*, conceituado como o processo que envolve o desenvolvimento de ferramentas e aplicação de técnicas para coletar, monitorar, analisar, resumir e visualizar conteúdos de mídias sociais (Zeng, Chen, Lusch & Li, 2010).

Em sua aplicação oferece oportunidades para descobrir e identificar padrões, características, informações e tópicos relevantes; originar perspectivas interessantes para a compreensão do comportamento humano; realizar análises qualitativas e quantitativas e até mesmo prever eventos futuros a partir de conteúdos não estruturados requerendo, para tanto, acompanhamento e refinamento de informações contínuo para alcançar bons resultados (Byun, Lee, You & Kim, 2013; Zeng, Li & Duan, 2012).

Lau, Lee e Ho (2005), ao tratarem especificamente da atividade turística, apresentam que as organizações turísticas públicas e privadas (OTPP) podem utilizar as informações adquiridas nas mídias sociais para elaborar novas estratégias, tomar decisões operacionais e elaborar novos produtos e serviços. Além disso, os autores acreditam que essa "inteligência competitiva" pode ajudar organizações a identificar vantagens, fraquezas, oportunidades, alavancar efetividade e melhorar a satisfação dos consumidores.

Para Carson (2008) as mídias sociais podem ser ferramentas extremamente eficazes para monitorar comentários e avaliações dos visitantes sobre sua

experiência no destino, compreender e conhecer as opiniões dos consumidores sobre os destinos, identificar forças e fraquezas do plano de marketing e das estratégias de promoção turística, bem como identificar e conhecer a imagem criada e percebida do destino na mente dos consumidores.

Ao monitorar a imagem da marca do destino tendo como fonte o conteúdo de usuários em mídias sociais, é possível medir a dissonância entre expectativa e realidade, abrindo espaço para identificação de pontos de desenvolvimento necessários. No entanto, Byun, Lee, You e Kim (2013) reconhecem a dificuldade de fazê-lo devido às principais características dos dados das mídias sociais: grandes, barulhentos e dinâmicos (*large, noisy, and dynamic*). Han, Kampler e Pei (2012) reforçam essa opinião ao lembrar que no SMM é necessário lidar com redes onipresentes de estruturas complexas.

3. Desafios na mineração de conteúdos em mídias sociais

Diante do grande volume e da natureza dinâmica dos conteúdos em mídias sociais gerados continuamente, coletar e identificar automaticamente temas emergentes e de interesse em meio a vibração das constantes conversas e interações entre usuários é apontado por diversos autores como um dos principais desafios no processo SMM. Parte significativa do conteúdo é elaborada ignorando regras de ortografia e gramática; apresentando problemas léxicos e sintáticos, como gírias, abreviações, ajustes de palavras, uso de *emoticons*, criação de novas palavras, significados múltiplos, entre outros (Abrahams, Jiao, Fan, Wang & Zhang, 2013; Paine, 2011).

Paine (2011) ressalta a coleta de conteúdos como a fase mais desafiadora do processo haja vista que serviços e *software* de SMM não oferecem garantia da integridade dos dados. Em média, 85% a 95% do conteúdo total existente de será coletado e, aproximadamente 70% é irrelevante. Portanto, compreende-se que o importante não seja o quanto está sendo coletado, mas sim a relevância de determinada amostra para os objetivos propostos.

Apesar de diversas ferramentas e *sites* permitirem a execução de fases do SMM, na maioria dos casos, a demanda por informação estratégica tem tamanho grau de especificidade que a generalização de padrões disponíveis não é suficiente, demandando o desenvolvimento de sistemas personalizados (Crooks, Croitoru, Stefanidis & Radzikowski, 2013). Portanto, as técnicas e ferramentas do SMM para coletar, compartilhar, investigar e visualizar dados de mídias sociais têm sido amplamente exploradas e desenvolvidas (Tang & Yang, 2012).

Seguindo o entendimento de Paine (2011) que até então não existia uma ferramenta, técnica ou metodologia confiável única para medir, monitorar, avaliar e analisar os conteúdos coletados, a combinação de diferentes técnicas de mineração pode ser o caminho para uma solução adequada. Han, Kamber e Pei (2012) afirmam que o desenvolvimento de métodos eficazes de descoberta de conhecimento e aplicações para um grande número de dados da rede é essencial e apesar de grande progresso ter sido feito, ainda há muitas questões em aberto para serem resolvidas. Abdel-Hafez e Xu (2013) destacam que, recentemente, esta área de pesquisa tem sido o foco de muitos pesquisadores e, portanto, os métodos propostos estão aumentando muito rapidamente.

4. Desenvolvimento de métodos de mineração de conteúdos em mídias sociais

No âmbito acadêmico, vários métodos para navegar, coletar, analisar o conteúdo, sentimentos e tópicos em mídias sociais têm sido propostos. Muitas técnicas de categorização de dados não estruturados foram adaptadas e aplicadas em mídias sociais em estudos de diversas áreas do conhecimento. Estes esforços de pesquisa exigem habilidades interdisci-

plinares, pois envolvem desde o trato de dados em bruto quanto a busca por informações de qualidade e com significado, exibidas da maneira mais adequada à situação (Abrahams et al., 2012).

Nikolov (2012) destaca que devido ao volume e complexidade dos dados, os modelos simples são inadequados e, portanto, o processo de exploração de dados de mídias sociais exige uma abordagem abrangente, destacando a necessidade de uma estrutura unificada para explorar a estrutura dos dados com eficiência para realizar a identificação, classificação e previsão de tópicos, assuntos e padrões relevantes.

Inúmeras técnicas como extração, classificação ou categorização, análise de características, análise linguística, análise de conteúdo, associação entre textos, agrupamento ou *clustering* e sumarização podem ser utilizadas para se extrair padrões ou conhecimentos interessantes e inesperados em documentos textuais (Bastos, 2006).

Para Chen e Liu (2004), a técnica de *clustering* ou agrupamento oferece a vantagem de revelar tendências imprevistas, correlações, ou padrões na estrutura dos dados que não haviam sido pressupostas. Lin, Hsieh e Chuang (2009) destacam que a análise de *clusters* é uma técnica bem explorada em mineração de dados e, segundo Abdous, He e Yen (2012), essa técnica exploratória permite visualizar padrões agrupando palavras e termos similares ou mesmo valores de atribuídos que são codificados de forma semelhante. Os mesmos autores afirmam ainda que sob uma perspectiva de mineração de dados, *clustering* é a descoberta não supervisionada de um padrão de dados ocultos e que esta abordagem é utilizada em situações em que um conjunto de registros pré-classificados está disponível.

Dentre os modelos identificados na literatura, o Processo de Descoberta de Conhecimento em Banco de Dados (KDD) proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) merece destaque pelo pioneirismo, aceitação e uso reconhecido por outros pesquisadores. O modelo sugerido por Han, Kamber e Pei (2012) destaca-se pela grau de atualização e pela notoriedade dos autores na área de mineração de dados (Chapman*;* Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2000). Já o modelo *Cross-Industry Standard Process for Data Mining* (CRISP-DM) destaca-se por ser considerado uma das metodologias mais populares e completas para aumentar o sucesso de projetos de mineração de dados, apresentada em muitas publicações da área e utilizadas na prática e por ser considerado o padrão de maior aceitação atualmente. As etapas e fases dos modelos descritos acima são apresentadas no quadro 1.

Em relação aos SMM destacam-se as metodologias propostas por Kalampokis, Tambouris e Tarabanis (2013), Abrahams, Jiao, Fan, Wang e Zhang (2013), He, Zha e Li (2013) e Neves (2013), apresentadas no quadro 2.

Apesar dos inúmeros métodos e técnicas, independente da mídia social investigada, a mineração de dados aplicada pode exigir abordagens únicas.

Quadro 1 | Modelos de Descoberta de Conhecimento em Bancos de Dados (KDD).

NOME	ETAPAS E FASES	AUTORES
Processo de descoberta de conhecimento em banco de dados (KDD)	Composto por cino etapas: (1) seleção dos dados; (2) pré-processamento e limpeza dos dados; (3) transformação dos dados; (4) mineração de dados; (5) interpretação e avaliação dos resultados.	Fayyad, Piatetsky-Shapiro e Smyth (1996).
Processo de descoberta de conhecimento em banco de dados (KDD)	Composto por 7 etapas: (1) limpeza de dados; (2) integração de dados; (3) seleção de dados; (4) transformação de dados; (5) mineração de dados; (6) avaliação de padrões; (7) apresentação do conhecimento.	Han, Kamber e Pei (2012)
Cross-industry standard process for data mining (CRISP-DM)	Composto por 5 etapas: (1) entendimento do negócio; (2) entendimento e compreensão dos dados; (3) preparação dos dados; (4) modelagem; (5) avaliação, utilização ou aplicação	Chapman et al., (2000)

Fonte: Elaboração própria.

Quadro 2 | Metodologias de *Social Media Mining*.

Fontes	ETAPAS E FASES
Kalampokis, Tambouris e Tarabanis (2013)	Composto por duas fases: fase de condicionamento de dados (coleta e filtragem de dados de mídias sociais brutos e computação de variáveis preditivas); fase de análise preditiva (criação do modelo de variáveis preditivas e avaliação do desempenho preditivo).
Abrahams, Jiao, Fan, Wang e Zhang (2013)	(1) Seleção de um fórum de discussão; (2) identificação e coleta de conteúdos em fóruns de discussão e mídias sociais; (3) seleção dos termos empregados e categorias determinantes dos termos mais expressivos; (4) armazenamento dos conteúdos e termos selecionados em base de dados; (5) mineração de texto; (6) classificação automática de componentes; (7) listagem de termos significativos por componente e categorias; (8) análise.
He, Zha e Li (2013)	Composto por três fases: (1) pré-processamento de texto (extração, preparação e coleta de conteúdos); (2) processamento e análise (coleta de conteúdo, aplicação de técnicas de mineração de texto como extração, categorização e agrupamento, resultados; (3) actionable intelligence (visualização dos resultados para identificar padrões, questões, tendências e modelos; recomendações e ações).
Neves (2013)	Composto por duas fases: (1) pré-análise (coleta, validação e seleção dos conteúdos); (2) exploração do material (identificação, classificação, direcionamento e normalização dos conteúdos).

Fonte: Elaboração própria.

Segundo Barbier e Liu (2011), diferentes conjuntos de dados e questões requerem diferentes tipos de ferramentas, ressaltando que o problema em si pode determinar a melhor abordagem. Ao compreendê-lo, pesquisadores e analistas devem selecionar uma abordagem de mineração de dados apropriada, realizando obrigatoriamente uma etapa de pré-processamento, visando eliminar informações irrelevantes ou fora do escopo.

5. Metodologia

Para o desenvolvimento desse estudo exploratório com técnica bibliográfica, foi realizado um levantamento de teses, dissertações e artigos científicos nas áreas de Ciência e Gestão da Informação, Sistemas de Informação, Tecnologia e Internet, Ciência da Computação, Recuperação de Informação e Mineração de Dados. O período de publicação das obras foi entre os anos de 2008 e 2013, tendo elas a obrigatoriedade da presença de um ou mais dos seguintes descritores: 'monitoramento de mídias sociais', 'mineração de dados', '*social media monitoring*', '*social media mining*', '*social media content mining*' e '*data mining*'.

A seleção da bibliografia de referência foi realizada por meio do Portal de Periódicos CAPES[2] [http://www.periodicos.capes.gov.br] e do sistema *Science Direct* [http://www.sciencedirect.com][3], duas bases de dados que indexam os principais periódicos nacionais e internacionais das áreas alvo. O quadro 3 apresenta os principais periódicos utilizados, estrato/conceito e área.

Após os conceitos reunidos na pesquisa bibliográfica, o desenvolvimento da metodologia foi pautado em três modelos de descoberta de conhecimento além de quatro SMM, apresentados em tópicos anteriores. Os três modelos de descoberta de conhecimento analisados foram os modelos propostos por Fayyad, Piatetsky-Shapiro e Smyth (1996), Han, Kamber e Pei (2012) e Chapman et al. (2000). Já as SMMs analisadas foram propostas por Kalampokis, Tambouris e Tarabanis (2013), Abrahams, Jiao, Fan, Wang e Zhang (2013), He, Zha e Li (2013) e Neves (2013).

Os modelos e metodologias analisadas foram selecionadas por apresentarem o processo de

[2] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, órgão público vinculado ao Ministério da Educação do Brasil.

[3] Cabe ressaltar que o acesso foi feito no âmbito da Universidade Federal do Paraná a qual, por estar vinculada à CAPES, possibilita acesso aos textos completos de artigos encontrados na busca.

construção metodológica e etapas da metodologia proposta de maneira clara, objetiva e detalhada, sendo fundamental para uma melhor análise e compreensão das etapas, fases e processos investigados. As etapas e fases analisadas foram combinadas e serviram de base para a construção da metodologia de SMM voltada à gestão de destinos turísticos proposta no presente estudo.

6. Resultados e discussão

Apesar dos modelos e metodologias selecionados terem aplicações, contextos, problemas e objetivos distintos, as principais fases e etapas mais importantes que compõem cada processo de SMM foram extraídas e combinadas, embasando a metodologia proposta. O processo da construção metodológica é composto por sete fases e será detalhado a seguir.

6.1. Fase I – Definição dos objetivos e preparação da coleta

A primeira fase deste processo tem por objetivo definir os objetivos e preparar a coleta dos conteúdos, consistindo em três etapas elaboradas a partir do modelo proposto por Kalampokis, Tambouris e Tarabanis (2013), para os quais a relevância dos dados depende de quatro perguntas: *when* (quando), *where* (onde), *who* (quem) e *what* (o que). As três etapas detalhadas que envolvem esta fase são descritas no quadro 4.

A etapa de identificação das características do perfil do usuário – relacionada ao perfil dos usuários – e a pergunta *who* (quem) não foram incluídas, pois

Quadro 3 | Lista dos periódicos utilizados.

Periódicos e Revistas	Conceito[4]	Área
Computers and Education	A1	Ciência da Computação
Communications in Computer and Information Science	B4	Ciência da Computação
Decision Support Systems	A1	Ciência da Computação
Enterprise Information Systems (Print)	B2	Ciência da Computação
Information Technology and Management	B1	Ciência da Computação
IEEE Intelligent Systems	A1	Ciência da Computação
International Journal of Multimedia Information Retrieval	B4	Ciência da Computação
International Journal of Information Management	A1	Administração, Ciências Contábeis e Turismo
Internet Research	B1	Ciência da Computação
International Journal of Computer Information Systems and Industrial Management Applications	B5	Ciência da Computação
Transactions in GIS (Print)	B2	Ciência da Computação
Tourism Management (1982)	B1	Interdisciplinar
The AI Magazine	A2	Ciência da Computação

Fonte: Elaboração própria.

Quadro 4 | Etapas da fase de coleta e filtragem de dados em bruto.

ETAPAS	DESCRIÇÃO
Determinação do período	Está relacionada com a questão *when* (quando), uma vez que especifica a duração e relação do período da coleta.
Localização	Está relacionado com a questão *where* (onde) os dados serão coletados.
Seleção de termos de busca	Está relacionada a seleção dos termos de pesquisa e corresponde a pergunta *what* (o quê).

Fonte: Kalampokis, Tambouris e Tarabanis (2013).

o presente método tem como objeto o conteúdo e não o usuário e suas características singulares.

Em relação às etapas de localização, definiu-se a mineração de conteúdos publicados pelos usuários sobre as cidades sedes da Copa do Mundo FIFA (Rio de Janeiro, São Paulo, Belo Horizonte, Porto Alegre, Brasília, Cuiabá, Curitiba, Fortaleza, Manaus, Natal, Recife e Salvador) e Foz do Iguaçu-PR, nas mídias sociais Facebook, Twitter e YouTube. A cidade de Foz do Iguaçu também foi selecionada por ser um dos principais destinos turísticos brasileiros visitados por estrangeiros e por ter sido selecionada como um dos destinos turísticos indutores pelo Ministério do Turismo.

Definiram-se quatro janelas temporais prévias ou contemporâneas à Copa do Mundo FIFA 2014 totalizando oito meses. O primeiro período de coleta atendeu à realização de testes e ocorreu durante a Copa das Confederações nos meses de julho, agosto e outubro de 2013. O segundo foi realizado em dezembro de 2013 em função do sorteio dos países, grupos e cidades-sede da Copa do Mundo FIFA 2014. O terceiro tem previsão para fevereiro, abril e maio de 2014 por ser o período que antecede o evento (Pré-Copa). O último período de coleta acontecerá em junho e julho de 2014, período de realização da Copa do Mundo FIFA 2014.

Já na terceira etapa, para selecionar os termos de pesquisa, no início de junho de 2013 foi construída uma ontologia de domínio partindo do estudo desenvolvido por Neves (2013) em similar aporte, porém voltado para os jogos olímpicos de Londres em 2012. O autor (2013) partiu de cinco categorias e 40 termos em inglês: alimentação (14 termos), hospedagem (8 termos), transportes (12 termos) e segurança (6 termos). Com enfoque em menções na mídia social Twitter, buscou mensagens onde havia uma associação obrigatória entre citação explícita do evento e cada um dos termos.

Algumas adaptações foram necessárias ao modelo de Neves (2013) para atender a demanda de monitoramento desse estudo. A categoria 'segurança' foi excluída; o escopo de termos foi ampliado para três idiomas: português, inglês e espanhol e; a quantidade de termos foi revista de forma a expandir algumas temáticas deixadas pelo autor por uma opção metodológica. Ainda, dada a importância que as OTPP têm para os objetivos, foram direcionados esforços para o levantamento dos principais atrativos turísticos das cidades-sede da Copa de 2014. No quadro 5 é possível verificar uma amostra dessa ontologia criada.

Quadro 5 | Amostra da ontologia de domínio desenvolvida para a pesquisa.

SECÇÃO	DIVISÃO	UNIDADE SEMÂNTICA	TERMO
serviços	hospedagem	hotel	hotel
			Hotels
			hoteles
			hotéis
...	...	...	...
	a&b	bar	bar
			pub
			beer

Fonte: Elaboração própria.

No quadro 5 a *sessão* é apenas uma denominação global que separa os serviços – correlatos ao estudo de Neves (2013) – dos atrativos, acrescentados nessa pesquisa. A *divisão* é a categoria, utilizada para futura análise agrupada dos resultados – similar ao *clustering*. A *unidade semântica* (US) é o objeto de nível de significado mais específico dentro das premissas do estudo, enquanto o *termo* é(são) a(s) sequência(s) de caracteres que será(ão) inserida(s) no sistema para que associações a uma US sejam encontradas. Por exemplo, os *termos* bar, *pub* e *beer* serão coletados, mas a todos eles será atribuído o significado de 'bar' para efeitos de análise. Quantos aos atrativos das cidades-sede, foi estabelecido o seguinte padrão na ontologia: a *divisão* corresponde à cidade; a US é o atrativo e os termos – analogamente ao supracitado – são as sequências de caracteres que levam até a US.

[4] Classificação "Qualis" de responsabilidade da CAPES.

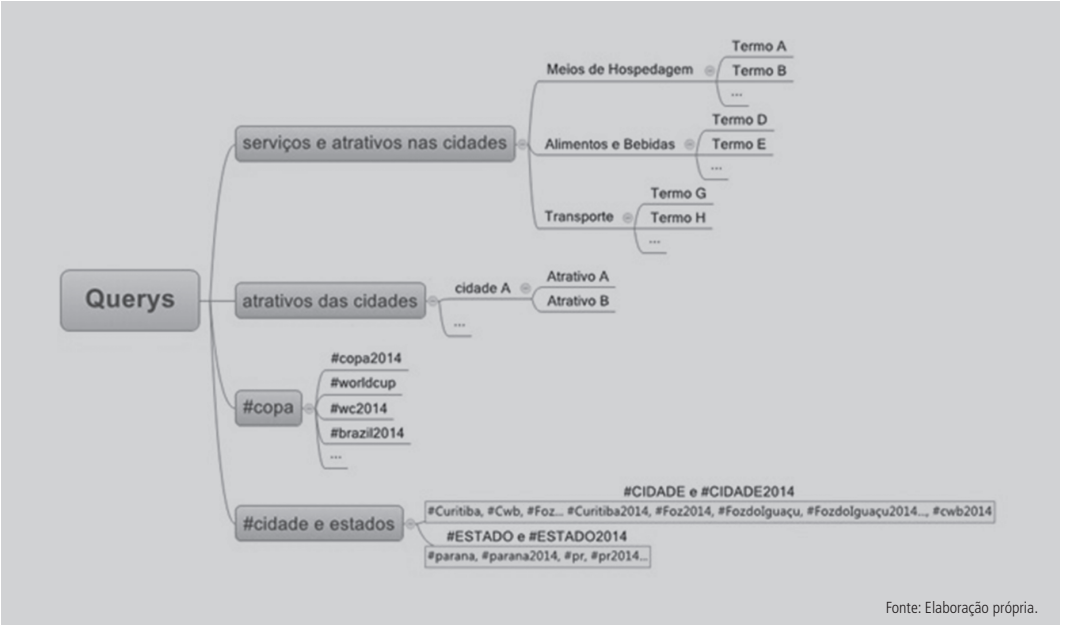


Figura 1 | Trabalhos (*querys*) que representam as estratégias de coleta de dados.

Conforme apresentado por Kalampokis, Tambouris e Tarabanis (2013), as diferentes abordagens para esta etapa dividem-se em categorias de seleção manual e dinâmica. Como coletar e codificar uma grande quantidade de dados em mídias sociais manualmente é tedioso e demanda muito tempo (He, Zha & Li, 2013), o presente estudo adota a abordagem dinâmica, onde os termos de pesquisa são obtidos através de um processo computacional. Sendo assim, a segunda fase consistiu na definição das ferramentas e *software* de coleta de dados em mídias sociais.

6.2. Fase II – definição das ferramentas e software de coleta

Primeiramente utilizou-se o *software* de monitoramento ACEITA, adquirido através de verba de projeto de pesquisa financiado pelo CNPq. No início do mês de Julho de 2013 e durante a primeira semana de Julho de 2013 foi realizado um pré-teste com o objetivo de conhecer e entender melhor sua operacionalização da coleta, funcionalidades e características.

Salvo as limitações número de palavras-chave/buscas a serem monitoradas (apenas 115), o pré-teste foi positivo. Porém, no período da Copa das Confederações o *software* apresentou problemas técnicos que não foram solucionados a tempo pelo suporte do monitoramento dos resultados. Após identificação e não solução dos problemas pela empresa foi encerrada a licença do *software*.

Para a melhor definição de uma nova ferramenta foi elaborada uma pesquisa comparativa da oferta nesse segmento, avaliando essencialmente: mídias sociais monitoráveis, carga de termos, carga de casos/menções, idiomas, custos, formato de exportação, possibilidade de recuperar dados anteriores, pré-processamento de dados e histórico da empresa. O sistema selecionado foi o Seekr Monitor, atendendo mais adequadamente às necessidades de configuração e coleta demandadas.

6.3. Fase III – Coleta

A fase de coleta de dados, consiste em três etapas, sendo a primeira etapa a elaboração da estratégia de coleta de conteúdo. No presente estudo a estratégia

adotada foi dividir a coleta em diversos trabalhos, sinônimos para consultas aos dados das mídias sociais. Essa divisão é necessária pela limitação técnica dos sistemas em relação às complexas e extensas solicitações (*querys*) demandadas. Aliando essa dificuldade à necessidade de categorizar os resultados, chegou-se num total de quatro trabalhos independentes representados por caixas paralelas na figura 1.

A primeira *query* refere-se a identificação de menções sobre a dimensão ontológica de serviços. A coleta inicial é feita pelo nome das cidades. Dos dados recuperados, são excluídos todos os casos em que não existe pelos menos um dos termos da ontologia (representados por Termo A, Termo B… na figura 1). O procedimento com atrativos é análogo: busca pelas cidades-sede e exclusão de menções que não contém os termos que representam os atrativos. Dessa maneira, há recuperação de uma grande quantidade de dados e no instante seguinte esse *corpus* é reduzido significativamente. Essa opção permitirá que o controle do conteúdo relevante seja feito pela pesquisador e não pela rede ou mídia social, o que aumenta a confiabilidade.

Uma abordagem diferente será feita com as *hashtags*, que é a adição do símbolo de *hash* (#) seguido de uma palavra, aplicada quando o usuário pretende associar sua publicação a um determinado tópico ou tema de discussão (Kwak, Lee, Park & Moon, 2010). Funcionam como etiquetas ou palavras-chave atribuídas a fotos, *posts*, vídeos, entre outros conteúdos publicados e compartilhados, que permitem ao usuário iniciar, encontrar e reunir conteúdos relevantes e conversas associadas a determinado tópico ou assunto com mais facilidade.

A vantagem em relação às duas *querys* anteriores é que a chance de coletar um conteúdo ambíguo será reduzida. Em geral, as combinações representam univocamente o evento e sua relação com cada cidade e estado. O objetivo dessas *querys* nesse estudo é contingencial, valendo-se da premissa que os usuários que se manifestam nas mídias sociais costumam usar as *hashtags* no caso desses megaeventos, conforme relatado por Paine (2011) e Neves (2013).

As *hashtags* atuam como uma alternativa para auxiliar no monitoramento, coleta e acompanhamento das referências ao evento, assunto e conversas em mídias sociais monitoradas em tempo real. No entanto, essa *query* é contingencial, pois apesar da alta popularidade e das vantagens da marcação social e utilização de *hashtags*, segundo Li e Lew (2012) as *tags* fornecidas pelos usuários podem ser inconsistentes e incompletas.

Assim, foram criados trabalhos para coletar as *hashtags* sobre o evento copa do mundo e especificamente sobre as cidades-sede. Cabe ressaltar que as *querys* poderão apresentar sobreposições e, por este motivo, demandarão um esforço posterior de eliminação de duplicatas (desde que todas as variáveis sejam iguais). Se uma mesma mensagem for emitida por mais de um usuário, o caso será mantido. Se um mesmo usuário emitir duas vezes a mensagem, um dos casos será excluído.

A segunda etapa da fase III foi extraída da metodologia CRISP-DM e consiste no entendimento e compreensão dos dados. Tem início com a coleta e prossegue com as atividades de aproximação e familiarização com a base, identificação de inconsistências, descoberta das primeiras características ou padrões nos dados e detecção de subconjuntos para formação de hipóteses sobre o conhecimento a ser descoberto (Chapman et al., 2000).

Já a terceira etapa corresponde ao pré-processamento de dados e consiste na transformação dos dados ruidosos (*noisy*) e brutos (*raw*), em formato utilizável, principalmente em relação a atribuição de caraterísticas e integração de dados (He, Zha & Li, 2013; Abrahams et al., 2013; Kalampokis, Tambouris & Tarabanis, 2013). Sendo assim, esta etapa consiste na classificação, exclusão, redução e armazenamento dos conteúdos recuperados, realizados de maneira automática através das funcionalidades e configurações do sistema Seekr Monitor.

Conforme indicado nos modelos de He, Zha e Li (2013) e Neves (2013), a quarta etapa da fase de coleta consiste em exportar os dados para planilhas eletrônicas para posterior limpeza e análise, correspondente à fase seguinte.

6.4. Fase IV – Tratamento e avaliação

Esta fase abrange todas as atividades e esforços do processo de descarte de dados inválidos e melhoramento dos restantes para construção se uma base de dados consolidada. Nesta etapa, os resultados ambíguos, inconsistentes e confusos serão tratados e as estratégias para resolver problemas nos dados serão estabelecidas. Como em toda análise quantitativa e qualitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis (He, Zha & Li, 2013; Abrahams et al., 2013; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Chapman et al., 2000).

O processo se iniciará com a (i) identificação de um padrão, pelo qual verifica-se a existência – frequente – de ocorrências que não atendem ao objetivo; (ii) uma solução é testada para que todos os casos que atendem ao padrão sejam separados; (iii) uma solução definitiva é executada. Para exemplificar a situação:

> **Padrão:** Manifestações religiosas
> **Descrição:** A palavra "Salvador" que dá nome ao município também tem um sentido religioso. Foram encontradas citações bíblicas ou menções ao "Salvador" enquanto figura religiosa.
> **Solução:** Excluir os casos em que Salvador não foi tratado como o nome do município.
> **Técnica:** Todos os resultados que contenham os seguintes termos foram excluídos: *evangelho evangélico evangelista evangélica jesus bíblia oração orar pastor gospel pastores Jehova "nosso salvador" culto "palavra de Deus" "nome de jesus" "palavra de cristo" "palabra de dios" "espírito santo" "tu és" oración*

Outros exemplos de ambiguidade neste objeto de estudo foram "Terra Natal" – cidade *versus* expressão; #fifa – o evento *versus* o lançamento do jogo.

Baseada na metodologia CRISP-DM, essa etapa consiste na avaliação e revisão das fases e etapas executadas anteriormente para se certificar que o modelo e os dados estão adequados aos objetivos definidos. (Chapman et al., 2000). Ela torna-se constante e exige revisão sistemática dos resultados para que um novo ciclo de identificação de padrões seja efetuado. Esclarece-se que padrões emergentes externos à ontologia do domínio, mas com valor para a pesquisa, também podem ser encontrados. Um exemplo foi a intensa polêmica sobre o valor dos ingresso cobrados para assistir aos jogos da Copa do Mundo FIFA 2014.

Após as fases de tratamento e avaliação, a fase V consiste na aplicação de técnica de mineração de dados.

6.5. Fase V – Mineração de dados

A fase de mineração de dados foi elaborada conforme os modelos de KDD propostos por Fayyad, Piatetsky-Shapiro e Smyth (1996), Han, Kamber e Pei (2012), CRISP-DM e nas metodologias propostas por Abrahams et al. (2013) e He, Zha e Li (2013).

Cabe ressaltar que as Fases de V a VII ainda não foram efetivamente testadas até a publicação deste artigo, ao contrário das anteriores. Nesse sentido, é a apresentada a metodologia prevista para aplicação.

Para o presente estudo, será utilizada a análise de *clusters* ou agrupamento, técnica bem explorada em mineração de dados para visualizar padrões agrupando palavras e termos similares ou valores atribuídos que são codificados de forma semelhante e permitem revelar tendências imprevistas, correlações, ou padrões na estrutura dos dados que não haviam sido pressupostas (Abdous, He & Yen, 2012; Lin, Hsieh & Chuang, 2009; Chen & Liu, 2004).

6.6. Fase VI - Identificação e interpretação de padrões

Com base em todos os modelos referência, a fase consistirá no pós-processamento, interpretação e análise dos padrões descobertos e extraídos, bem como a possibilidade de detalhamento em estudos comparativos de padrões, tópicos, assuntos,

relações, conhecimentos e tendências. Para análise dos conteúdos obtidos será utilizada um *software* especializado no suporte à análise de conteúdo, para onde será exportada a base de dados trabalhada até então. Após as fases de identificação e interpretação de padrões, a fase VII consiste na apresentação dos principais padrões identificados nos dados.

6.7. Fase VII - apresentação de tendências e principais padrões encontrados

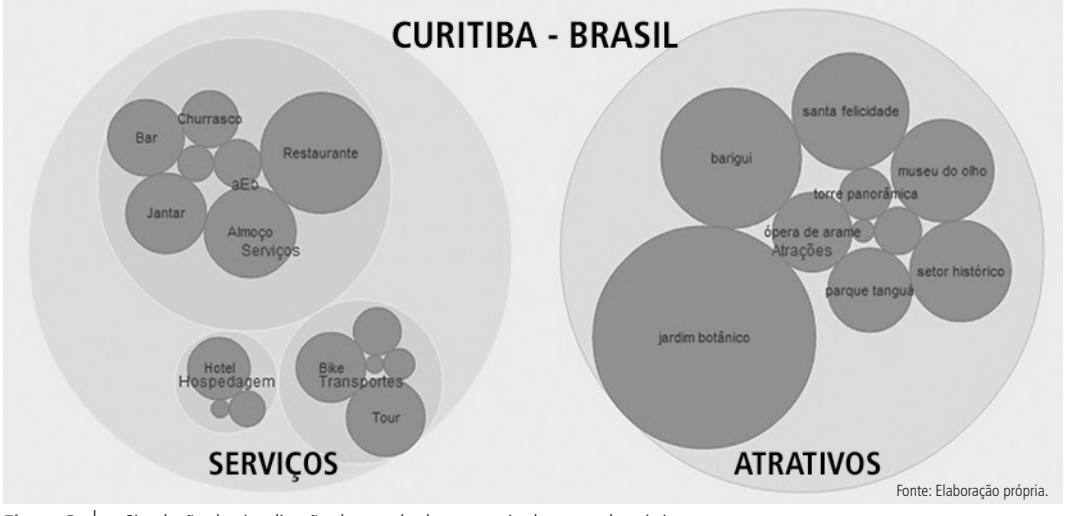
Elaborada a partir do modelo proposto por Han, Kampler e Pei (2012), nesta etapa a base de dados final será transformada e representada utilizando técnicas de visualização e exibição. Para o presente estudo, o modelo de apresentação dos principais padrões nos dados será feita através da criação de mapas heurísticos, conforme simulação apresentados na figura 2.

Para a exploração do aspectos visual outro software dedicado será utilizado, tendo sido desenvolvido especificamente para o estudo em questão pela Aquarela, financiado por recursos de projeto do Conselho Nacional de Desenvolvimento Científico (CNPq). O *software* permitirá a concepção dinâmica dos mapas heurísticos, facilitando a visualização de padrões e relações encontrados nos conjuntos de dados em tempo real. Também permitirá explorar e realizar análises comparativas entre os resultados dos conteúdos dos destinos turísticos monitorados e investigados de acordo com os períodos e categorias definidas.

7. Considerações finais

Devido as suas características e funcionalidades, além da interatividade e comunicação, as mídias sociais também podem ser ferramentas extremamente eficazes e assumir um importante papel na gestão, no planejamento estratégico, desenvolvimento e fomento da atividade turística.

Para isso, por serem responsáveis pelo planejamento, gestão, promoção e desenvolvimento da atividade turística em seus respectivos destinos, é fundamental que as OTPP estejam atentas a este cenário de oportunidades que o ambiente das mídias sociais oferece. O conteúdo potencialmente coletado, monitorado e interpretado pode ser utilizado como instrumento de suporte à gestão, planejamento, desenvolvimento, processos de tomadas de decisão estratégicas e operacionais, criação de estra-



Fonte: Elaboração própria.

Figura 2 | Simulação da visualização dos resultados por meio de mapas heurísticos.

tégias de *marketing*, inovação, bem como oferecer oportunidades de melhoria e até mesmo criação de novos produtos e serviços turísticos.

É importante destacar que o SMM não deve ser resumidos em métricas e coleta de informações e experiências dos usuários. É preciso interpretá-las, analisá-las e estudá-las para a partir disso, obter informações e conhecimentos relevantes, identificar oportunidades, falhas, necessidades, expectativas, experiências, desejos, críticas, opiniões e avaliações dos usuários e consumidores (potenciais e reais) sobre os destinos turísticos.

Durante o processo de investigação, foi possível identificar que o SMM é um desafio e diversos modelos e métodos para navegar, coletar, analisar o conteúdo, sentimento e tópicos em mídias sociais têm sido propostos por pesquisadores de diversos campos do conhecimento.

Sendo assim, com o presente estudo contribuiu-se para criação e consolidação de um SMM voltado à gestão de destinos turísticos, composto por sete fases: (i) definição dos objetivos e preparação da coleta; (ii) definição das ferramentas e *softwares* de coleta; (iii) coleta; (iv) tratamento avaliação do conteúdo coletado; (v) mineração de dados; (vi) identificação e interpretação de padrões; e (vii) apresentação de tendências e principais padrões encontrados.

Ressalta-se que por ainda não ter sido concluída a aplicação ao caso da Copa do Mundo 2014, os parâmetros das etapas cinco, seis e sete podem ser alterados e adaptados. Pretendem-se, nesse sentido, voltar a ampliar essa discussão no âmbito da comunicação científica em fase posterior. Outros pesquisadores poderiam também testar essa metodologia em situações análogas para contribuir com seu melhoramento e consolidação no cenário do SMO aplicado aos destinos turísticos.

Referências bibliográficas

Abdel-Hafez, A., & Xu, Y. (2013). A survey of user modelling in social media websites. *Computer and Information Science, 6*(4), 59-71.

Abdous M. H., He, W., & Yen, C. J. (2012) Using data mining for predicting relationships between online question theme and final grade. *Educational Technology & Society, 15*(3), 77-88.

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2012). What's buzzing in the blizzard of buzz?: Automotive component isolation in social media postings. *Decision Support Systems, 55*(4), 871-882.

Barbier, G. & Liu, H. (2011). Data mining in social media. In C. Aggarwal (Ed.), *Social network data analytics* (pp. 327-352). New York: Springer.

Bastos, V. M. (2006). *Ambiente de descoberta de conhecimento na web para a língua portuguesa*. Tese de Doutoramento, UFRJ/ COPPE, Rio de Janeiro.

Blackshaw, P., & Nazzaro, M. (2006). Consumer-generated media (CGM) 101: Word-of-mouth in the age of the web-fortified consumer. *A Nielsen BuzzMetrics White Paper*. New York.

Byun, C., Lee, H., You, J., & Kim, Y (2013). Efficient keyword-related data collection in a social network with weighted seed selection. *International Journal of Networked and Distributed Computing, 1*(3), 167-173.

Carson, D. (2008). The blogshere as a market research tool for tourism destination: A case study of Australia's northern territory. *Journal of Vacation Marketing, 14*(2), 111-119.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS Inc.

Chen, S. Y., & Liu, X. (2004). The contribution of data mining to information science. *Journal of Information Science, 30*(6), 550-558.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*(1), 124-147.

Dai, Y., Kakkonen, T., & Sutinen, E. (2011). MinEDec: A decision-support model that combines text-mining technologies with two competitive intelligence analysis methods. *International Journal of Computer Information Systems and Industrial Management Applications, 3*, 165-173.

Dey, L., Haque, S. M., Khurdiya, A., & Shroff, G. (2011, September 17). Acquiring competitive intelligence from social media. *Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data*, Beijing.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37-54.

Governatori, G., & Iannella, R. (2011). A modelling and reasoning framework for social networks policies. *Enterprise Information Systems, 5*(1), 145-167.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Waltham: Elsevier.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464-472.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research, 23*(5), 544-559.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April 26-30). What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 591-600). Raleigh.

Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. *Cornell Hotel and Restaurant Administration Quarterly, 46*(3), 344-362.

Li, Z., & Lew, M. S. (2012). Cost-sensitive learning in social image tagging: review, new ideas and evaluation. *International Journal of Multimedia Information Retrieval, 1*(4), 205-222.

Lin, F. R., Hsieh, L. S., & Chuang, F. T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education, 52*(2), 481-495.

Nikolov, S. (2012). *Trend or no trend: A novel nonparametric method for classifying time series*. Tese de Doutoramento, Massachusetts Institute of Technology, Cambridge.

Neves, A. J. W. A. (2013). *Qualidade percebida de produtos e serviços turísticos em eventos*. Dissertação de Mestrado, Universidade Federal do Paraná, Curitiba.

Paine, K. D. (2011). *Measure what matters: Online tools for understanding customers, social media, engagement, and key relationships*. Westford: Wiley.

Tang, X., & Yang, C. C. (2012). Social network integration and analysis using a generalization and probabilistic approach for privacy preservation. *Security Informatics, 1*(1), 1-14.

Torres, C. (2009). *A bíblia do marketing digital: Tudo o que você precisa saber sobre marketing e publicidade na internet e não tinha a quem perguntar*. São Paulo: Novatec.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*(2), 179-188.

Zeng, L., Li, L., & Duan, L. (2012). Business intelligence in enterprise computing environment. *Information Technology & Management, 4*(13), 297-310.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *Intelligent Systems, IEEE, 25*(6), 13-16.