
**ToPI - Uma abordagem *online* para identificar
locais de interesse utilizando fotografias
geo-referenciadas**

Jean Roberto Ponciano



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2016

Jean Roberto Ponciano

**ToPI - Uma abordagem *online* para identificar
locais de interesse utilizando fotografias
geo-referenciadas**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Profa. Dra. Denise Guliato

Coorientador: Prof. Dr. José Gustavo de Souza Paiva

Uberlândia

2016

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

P795t Ponciano, Jean Roberto, 1990-
2016 ToPI - uma abordagem online para identificar locais de interesse
utilizando fotografias geo-referenciadas / Jean Roberto Ponciano. - 2016.
94 f. : il.

Orientadora: Denise Guliato.
Coorientador: José Gustavo de Souza Paiva.
Dissertação (mestrado) - Universidade Federal de Uberlândia,
Programa de Pós-Graduação em Ciência da Computação.
Inclui bibliografia.

1. Computação - Teses. 2. Banco de dados - Teses. 3. Fotografias -
Teses. 4. Turismo - Teses. I. Guliato, Denise. II. Paiva, José Gustavo de
Souza, 1979- III. Universidade Federal de Uberlândia, Programa de Pós-
Graduação em Ciência da Computação. IV. Título.

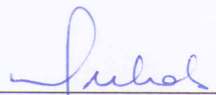
CDU: 681.3

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada "**ToPI - Uma abordagem online para identificar locais de interesse utilizando fotografias geo-referenciadas**" por **Jean Roberto Ponciano** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

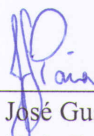
Uberlândia, 06 de abril de 2016

Orientador:



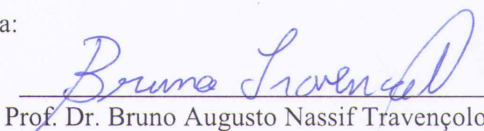
Prof.ª Dr.ª Denise Guliato
Universidade Federal de Uberlândia

Coorientador:

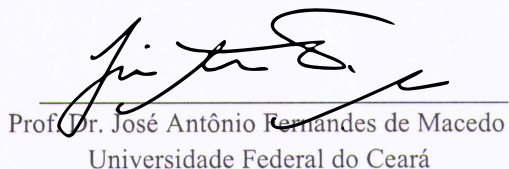


Prof. Dr. José Gustavo de Souza Paiva
Universidade Federal de Uberlândia

Banca Examinadora:



Prof. Dr. Bruno Augusto Nassif Travençolo
Universidade Federal de Uberlândia



Prof. Dr. José Antônio Fernandes de Macedo
Universidade Federal do Ceará

Aos meus pais Giovanni e Idalena e à minha irmã Adrielly.

Agradecimentos

A Deus, que me proporciona as oportunidades e a força de vontade para agarrá-las.

Aos meus pais Valter Giovani e Idalena Aparecida e à minha irmã Adrielly Cristina pelo apoio incondicional, conselhos, paciência e entendimento quanto às inúmeras ausências nos fins de semana e pedidos por silêncio. Sei que não foi fácil.

À minha orientadora Denise Guliato, pela orientação e pelo apoio e paciência que sempre teve comigo.

Ao meu coorientador José Gustavo de Souza Paiva, sempre comprometido, presente e participativo. Obrigado!

A todos os amigos que estiveram comigo nesta caminhada, em especial ao Caio, Cláudio e Everton, parceiros desde a graduação. Em breve alcançarei vocês no doutorado!

À CAPES, pelo suporte financeiro.

“Por mais longa que seja a caminhada, o mais importante é dar o primeiro passo.”
(Vinícius de Moraes)

Resumo

Devido ao crescente uso de redes sociais, as pessoas deixaram de ser apenas consumidores de informações, elas passaram a também produzi-las e compartilhá-las. Informações geo-referenciadas, isto é, informações com dados de localização geográfica, têm sido utilizadas em várias propostas da literatura para identificar locais de interesse e auxiliar turistas que visitarão cidades que ainda não lhe são familiares. Este trabalho apresenta uma estratégia *online* que utiliza fotografias geo-referenciadas e seus metadados para identificar locais de interesse pertencentes a uma dada região geográfica e recuperar informações relevantes relacionadas. Todo o processo é executado automaticamente e em tempo real, retornando informações atualizadas sobre os locais. A estratégia proposta leva em consideração o dinamismo inerente a redes sociais e, assim, é robusta quanto a inconsistências e/ou informações desatualizadas, problemas comuns em soluções que se baseiam em dados pré-armazenados. A análise de resultados mostrou que a proposta é bastante promissora, retornando locais que apresentam alta taxa de concordância em relação àqueles existentes em um *website* turístico bastante popular.

Palavras-chave: turismo, identificação de locais turísticos, contexto dinâmico, locais de interesse, fotografias, informações geo-referenciadas.

Abstract

Due to the growing use of social networks people no longer just consume data, they also produce and share it. Geo-tagged information, i.e., data with geographical location, have been used in many attempts to identify popular places and help tourists that will visit unfamiliar cities. This Master Thesis presents an online strategy that uses geo-tagged photos and their metadata in order to identify places of interest inside a given geographical area and retrieve relevant related information. The whole process runs automatically in real time, returning updated information about places. The proposed strategy takes into account the inherent dynamism of social media, and thus is robust under inconsistencies and/or outdated information, a common issue in solutions that rely on previously stored data. The analysis of the results showed that our approach is very promising, returning places that present high agreement with those from a popular travel website.

Keywords: tourism, identification of tourist places, dynamic context, places of interest, photographs, geo-tagged data.

Lista de ilustrações

Figura 1	– Exemplos de inconsistências: (a) Tela do sistema TripBuilder (BRILHANTE et al., 2013) - <i>screenshot</i> obtido em (BRILHANTE et al., 2014b); (b) Tela do sistema VIRaL (AVRITHIS et al., 2010a) - <i>screenshot</i> obtido em (AVRITHIS et al., 2010b) e adaptado pelo autor.	28
Figura 2	– Fotografia da Torre Eiffel e alguns de seus metadados. Fonte: Flickr	32
Figura 3	– Grupos com formatos variados. Imagem adaptada de (HAN; KAMBER; PEI, 2011).	33
Figura 4	– Exemplo de grupos com diferentes densidades. A utilização de valores únicos e globais de densidade resultaria na obtenção dos grupos A, B e C. No entanto, o grupo C possui subgrupos que não seriam identificados (C1, C2 e C3). Imagem adaptada de (ANKERST et al., 1999).	36
Figura 5	– Identificação dos grupos a partir da ordenação resultante. Imagem adaptada de (ANKERST et al., 1999).	38
Figura 6	– Grupo composto pelos elementos do intervalo [3, 16]. Imagem adaptada de (ANKERST et al., 1999).	38
Figura 7	– Sucessivos deslocamentos de média definem um caminho que leva à máxima densidade local. Imagem adaptada de (COMANICIU; MEER, 1999).	40
Figura 8	– Tela do sistema <i>World Explorer</i> exibindo partes de São Francisco, EUA. O usuário selecionou a <i>tag</i> “ <i>Golden Gate Bridge</i> ” (porção superior à esquerda) para que fossem exibidas imagens com essa <i>tag</i> nessa região geográfica. Imagem adaptada de (AHERN et al., 2007).	44
Figura 9	– ToPI - Uma visão geral do fluxo de dados.	48
Figura 10	– Gráfico de dispersão com 500 fotografias recuperadas nas proximidades do Cristo Redentor (representado pelo ponto destacado em azul no centro da imagem).	49

Figura 11 – Exemplo de união entre PoIs: (a) União automática entre os PoIs 1 e 2 (regra 1); (b) União entre os PoIs 3 e 4 apenas se existir palavra-chave coincidente nos conjuntos N_{pc}^3 e N_{pc}^4 (regra 2).	52
Figura 12 – Dois prédios chamados “ <i>Flatiron Building</i> ” em diferentes cidades dos Estados Unidos: (a) New York City, New York; (b) Fort Worth, Texas.	53
Figura 13 – Exemplo que destaca a importância de se priorizar a busca textual em detrimento da busca por proximidade geográfica na obtenção de informações de um PoI.	54
Figura 14 – Visão geral do sistema ToPI <i>Trip</i>	56
Figura 15 – Alguns Pontos de Interesse (PoIs) identificados a partir de uma busca em Paris, França.	57
Figura 16 – Níveis de <i>zoom</i> e visualização dos PoIs: (a) Nível de <i>zoom</i> mais distante do nível da rua; (b) Duas regiões e um PoI que aparecem ao dar <i>zoom</i> na região destacada pelo círculo vermelho em (a); (c) Os três PoIs pertencentes à região destacada em (b).	57
Figura 17 – Informações sobre a ponte do Brooklyn, Nova York, EUA: (a) informações básicas; (b) frequência de visitas realizadas ao longo de 2014 (funcionalidade acessível por meio do botão destacado pelo círculo vermelho em (a)).	58
Figura 18 – Exibição de PoIs no mapa conforme a popularidade mínima escolhida pelo usuário.	59
Figura 19 – Informações sobre a Torre Eiffel (Paris, França) retornadas e exibidas pelo <i>Bing Knowledge Widget</i>	60
Figura 20 – Exemplos de cidades onde encontram-se as localizações pesquisadas.	60
Figura 21 – Reconhecimento de PoIs pelo <i>Bing Knowledge Widget</i> : (a) PoI reconhecido - por isso há o sublinhado no nome; (b) PoI não reconhecido (sem sublinhado) - nesse caso, o usuário pode selecionar o nome e solicitar uma busca por meio do botão “ <i>bing</i> ”.	61
Figura 22 – Sequência de fluxos executados pelo sistema.	62
Figura 23 – Alguns PoIs retornados na busca em Paris.	70
Figura 24 – Situação “lugar vs evento” no qual o PoI foi identificado como <i>Paris Games Week</i> , um evento anual sediado no <i>Paris expo Porte de Versailles</i>	71
Figura 25 – PoI identificado incorretamente. O PoI destacado pelo círculo vermelho não é popular por si mesmo, mas sim um local onde as pessoas tiraram fotografias dos locais <i>Square Rene Viviani</i> e <i>Notre-Dame Cathedral</i> (também identificado como um PoI e representado pelo outro marcador na imagem).	71
Figura 26 – Identificação do PoI <i>Arc de Triomphe</i> sem tratativa de PoIs redundantes (a); com a eliminação de redundâncias (b).	72

Figura 27 – Alguns PoIs retornados na busca no Rio de Janeiro.	73
Figura 28 – Dois grupos representando o mesmo PoI (<i>Sugarloaf Mountain</i> ou Morro Pão de Açúcar, nome que aparece na imagem).	73
Figura 29 – Alguns PoIs retornados na busca em Nova York.	74
Figura 30 – Erro na obtenção de informações de um PoI: (a) PoI rotulado incorretamente (o correto seria <i>The Bushwick Collective</i>); (b) Fotografia pertencente ao PoI com o nome do local, reafirmando o erro na rotulagem.	75
Figura 31 – Situação “PoIs dentro de PoIs”: Uma busca retornou o edifício <i>Rockefeller Center</i> (imagem à esquerda) e outra retornou o <i>Rainbow Room</i> , restaurante dentro dele (imagem à direita).	76
Figura 32 – Situação de dinamicidade do repositório de fotografias: (a) Consulta por PoIs realizada utilizando fotos tiradas antes dos atentados terroristas de 2015 em Paris; (b) Consulta utilizando fotos tiradas em um intervalo de tempo que cobre os atentados e consequente identificação de dois locais atacados.	77
Figura 33 – Dinamicidade causada pela copa do Mundo 2014: (a) Consulta realizada após a realização do evento; (b) Consulta com fotos tiradas durante a copa, possibilitando a identificação dos dois locais apresentados.	78
Figura 34 – Locais muito movimentados durante a copa do mundo de 2014: (a) Consulta realizada após a realização do evento - sem fotos suficientes para identificar o estádio do Maracanã e o aeroporto Santos Dumont; (b) Consulta com fotos tiradas durante a copa, possibilitando a identificação desses dois locais.	79
Figura 35 – Exemplo de identificação de PoIs recentes: <i>Museu do Amanhã</i> (<i>Museum of Tomorrow</i>), inaugurado em 17 de dezembro de 2015 no Rio de Janeiro.	80
Figura 36 – Influência de um evento na identificação de PoIs: (a) Consulta realizada após a realização do GP de Mônaco 2015 - nenhum PoI identificado na região da imagem; (b) Consulta com fotos tiradas no período do evento, possibilitando a identificação de um PoI.	81

Lista de tabelas

Tabela 1	–	Identificação de palavras-chave. Exemplo 1.	51
Tabela 2	–	Identificação de palavras-chave. Exemplo 2.	51
Tabela 3	–	PoIs utilizados como <i>ground truth</i>	64
Tabela 4	–	Comparação dos métodos de agrupamento baseados em densidade utilizando <i>F-Measure</i> . Quanto mais próximo do valor 1 (um), melhor. . .	65
Tabela 5	–	Comparação dos métodos de agrupamento em relação ao tempo médio de execução (em segundos).	66
Tabela 6	–	Valores dos parâmetros utilizados nos experimentos.	68
Tabela 7	–	Cidades pesquisadas nos experimentos.	68
Tabela 8	–	Resultados dos experimentos. PC_x indica qual das quatro palavras-chave foi usada na identificação do PoI. C_x refere-se ao critério utilizado para que houvesse concordância.	69
Tabela 9	–	Comparativo entre a quantidade de tentativas gastas na obtenção de informações utilizando (1) palavras-chave e (2) as tags com as maiores pontuações.	82
Tabela 10	–	Tempo de execução de cada etapa do método ToPI (em segundos). . .	83

Lista de Algoritmos

1	DBSCAN	35
2	OPTICS	37
3	<i>Mean Shift</i>	39
4	HDBSCAN	41
5	P-DBSCAN	42

Lista de siglas

API Interface de Programação de Aplicativos - *Application Programming Interface*

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*

GPS Sistema de Posicionamento Global - *Global Positioning System*

HDBSCAN *Hierarchical DBSCAN*

MST Árvore geradora de custo mínimo - *Minimum Spanning Tree*

OMT Organização Mundial do Turismo

OPTICS *Ordering Points To Identify the Clustering Structure*

PoI Ponto de Interesse - *Point of Interest*

P-DBSCAN “Photo” DBSCAN

TF-IDF *Term Frequency-Inverse Document Frequency*

ToPI *Tourist Place Identification*

Sumário

1	INTRODUÇÃO	27
1.1	Objetivos e Hipótese	29
1.2	Contribuições	29
1.3	Organização da Dissertação	29
2	FUNDAMENTAÇÃO TEÓRICA	31
2.1	Conceitos básicos	31
2.1.1	Imagens com metadados	31
2.1.2	Agrupamento de dados	31
2.2	Trabalhos Correlatos	43
2.3	Considerações Finais	45
3	MÉTODO ToPI	47
3.1	ToPI	47
3.1.1	Seleção das fotografias	48
3.1.2	Agrupamento e filtragem	49
3.1.3	Identificação das palavras-chave que descrevem os PoIs	50
3.1.4	Tratativa de PoIs redundantes	51
3.1.5	Obtenção de informações sobre os PoIs	52
3.2	Considerações Finais	54
4	SISTEMA ToPI Trip	55
4.1	Visão geral do sistema	55
4.2	Funcionalidades do sistema ToPI <i>Trip</i>	56
4.3	Execução em tempo real	59
4.4	Configuração do sistema	61
4.4.1	Escolha do repositório <i>online</i> de fotografias	62
4.4.2	Método de pontuação das tags	63

4.4.3	Método de agrupamento	64
4.5	Considerações Finais	66
5	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	67
5.1	Relevância dos PoIs identificados	67
5.1.1	Método para a Avaliação	67
5.1.2	Experimentos	69
5.2	Identificação de PoIs em contexto dinâmico	76
5.2.1	Atentados terroristas em Paris - nov/2015	76
5.2.2	Copa do mundo de 2014	78
5.2.3	Museu do Amanhã - Rio de Janeiro	80
5.2.4	Grande prêmio de Mônaco de Fórmula 1 - 2015	80
5.3	Importância da utilização de palavras-chave	81
5.4	Tempo de execução	82
5.5	Considerações Finais	83
6	CONCLUSÃO	85
6.1	Limitações e Impactos	86
6.2	Trabalhos Futuros	86
	REFERÊNCIAS	89

Introdução

Segundo a Organização Mundial do Turismo (OMT), o turismo mundial segue uma tendência de crescimento. Em 2013, por exemplo, o turismo internacional alcançou o número recorde de 1,087 bilhão de chegadas, o que equivale a um aumento de 5% em relação a 2012. Em 2014 um novo recorde: 1,1 bilhão de chegadas, 4,7% a mais que 2013. A perspectiva também é positiva a longo prazo, com estimativa de crescimento anual de 3,3% entre 2010 e 2030 (UNWTO, 2015; UNWTO, 2014).

Parte das viagens turísticas internacionais, bem como parte das nacionais, é realizada por turistas que visitam seu destino pela primeira vez, o que exige deles um certo esforço para obter informações dos locais potencialmente interessantes presentes na(s) cidade(s) de destino, possibilitando assim a tomada de decisão sobre quais lugares visitar. Trata-se de um planejamento geralmente complexo e demorado que pode envolver a consulta a *websites* de turismo, plataformas públicas de informação, mapas turísticos, redes sociais, blogs pessoais, sugestões de amigos, entre outros.

Diversos trabalhos científicos têm sido desenvolvidos com o objetivo de auxiliar o turista na tarefa de encontrar locais de interesse para visitaç o. Algumas propostas, como (BRILHANTE et al., 2013), procuram ajudar turistas “de primeira viagem” recomendando roteiros de visitas es; outras, como (GAGGI, 2013), consideram que o usu rio j  possui um planejamento pr vio mas deseja conhecer locais pr ximos   rota escolhida (como, por exemplo, uma cachoeira pr ximo a uma rodovia); e outras ainda visam atender dois p blicos (LEMOS et al., 2012): aqueles que est o visitando o lugar pela primeira vez e aqueles que j  conhecem o local de interesse, mas procuram por novos pontos de vista dele.

Com o crescente uso de plataformas sociais e a populariza o de dispositivos com Sistema de Posicionamento Global - *Global Positioning System* (GPS), v rias propostas da literatura passaram a utilizar imagens com geo-localiza o e outras informa es textuais, obtidas em redes sociais, na elabora o de sistemas de recomenda o de locais tur sticos, sum rios visuais de lugares populares, sistemas de visualiza o e outros. Muitos trabalhos, no entanto, armazenam em bases de dados as informa es que s o

utilizadas posteriormente para a identificação dos Pontos de Interesse (PoIs) ou identificam esses locais uma única vez, armazenando-os para consultas posteriores. Por mais que exista uma rotina de atualização nessas bases, denominadas neste trabalho como bases estáticas, determinar o melhor intervalo de atualização não é trivial. Além disso, requer o armazenamento de uma quantidade de dados em constante crescimento e/ou uma seleção criteriosa sobre o que deve ser mantido e o que deve ser descartado, o que cria um esforço adicional.

Alguns autores, como Brilhante et al. (BRILHANTE et al., 2014a), defendem que a utilização de bases de dados com as informações necessárias torna a aplicação mais flexível, já que ela seria independente da fonte utilizada para obtenção dos PoIs, bastando consultar a base. No entanto, a utilização desse tipo de base de dados pode resultar em inconsistências, como as apresentadas na Figura 1, em que fotos presentes na base de dados não existem mais no repositório de fotografias. Mais importante que isso, aplicações com essa característica podem continuar apresentando PoIs que não existem mais ou deixar de retornar PoIs abertos recentemente, confundindo o usuário e prejudicando seu planejamento.

O presente trabalho apresenta um novo método, denominado *Tourist Place Identification* (ToPI), que visa identificar PoIs em contexto dinâmico, isto é, em um contexto no qual as informações disponíveis sobre PoIs podem sofrer alterações com o passar do tempo.

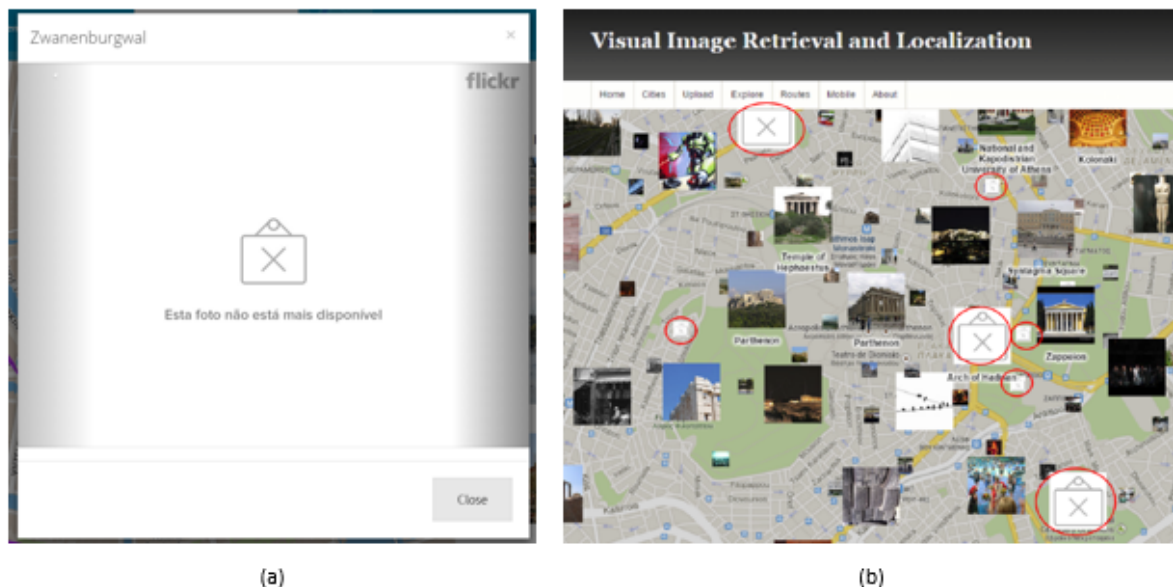


Figura 1 – Exemplos de inconsistências: (a) Tela do sistema TripBuilder (BRILHANTE et al., 2013) - *screenshot* obtido em (BRILHANTE et al., 2014b); (b) Tela do sistema VIRaL (AVRITHIS et al., 2010a) - *screenshot* obtido em (AVRITHIS et al., 2010b) e adaptado pelo autor.

1.1 Objetivos e Hipótese

O objetivo deste trabalho é identificar automaticamente e em tempo real pontos de interesse nas cercanias de uma dada localização, baseando-se em repositórios *online* de fotos geo-referenciadas. Os objetivos específicos, por sua vez, são:

1. Identificar e rotular pontos de interesse com base nos metadados de fotografias tiradas nas proximidades de uma localização dada pelo usuário;
2. Recuperar informações relevantes sobre cada ponto de interesse identificado;
3. Implementar um sistema de apoio à exploração geográfica de pontos de interesse.

A hipótese é que a identificação automática de pontos turísticos sob contexto dinâmico pode ser obtida a partir da aglomeração, em tempo real, de fotografias geo-referenciadas disponíveis em repositórios *online*.

1.2 Contribuições

As principais contribuições do trabalho são:

- ❑ Um método de identificação de PoIs, não-supervisionado e em tempo real, capaz de identificar pontos turísticos em qualquer localidade desde que haja fotografias com metadados associados em repositórios *online* para a localidade desejada;
- ❑ Um método de recuperação de informações sobre os PoIs que auxilie a decisão sobre visitas;
- ❑ Uma aplicação *Web* para auxiliar o usuário a encontrar e aprender sobre pontos turísticos próximos à localização desejada.

1.3 Organização da Dissertação

O restante da dissertação foi dividido em cinco capítulos, organizados da seguinte forma:

- ❑ Capítulo 2: Descreve sucintamente os principais conceitos básicos que norteiam o desenvolvimento do trabalho e o situa perante o estado da arte, apresentando o que tem sido feito na literatura e as limitações existentes em relação à identificação de pontos turísticos.
- ❑ Capítulo 3: Descreve o método ToPI, proposto neste trabalho de mestrado.

- ❑ Capítulo 4: Apresenta o sistema ToPI *Trip*, desenvolvido para validar o método proposto e ajudar o usuário final na exploração do conhecimento obtido.
- ❑ Capítulo 5: Descreve os experimentos realizados e analisa seus resultados. Apresenta também situações especiais, decorrentes, na maioria das vezes, da dinamicidade inerente às fontes de informações utilizadas.
- ❑ Capítulo 6: Encerra a dissertação, apresentando as conclusões obtidas e perspectivas de trabalhos futuros.

Fundamentação Teórica

Este capítulo é dividido em duas partes. A Seção 2.1 descreve sucintamente os conceitos teóricos necessários para o entendimento e desenvolvimento do trabalho. A Seção 2.2, por sua vez, situa a proposta deste trabalho em relação ao estado da arte, apresentando trabalhos encontrados na literatura a respeito da identificação de locais turísticos.

2.1 Conceitos básicos

2.1.1 Imagens com metadados

As imagens utilizadas pelas propostas da literatura que serão apresentadas na Seção 2.2, bem como pelo método desenvolvido neste trabalho, precisam estar associadas a informações textuais que as caracterizam, chamadas de metadados. Tais metadados serão utilizados na identificação e representação dos pontos turísticos retornados pelas propostas.

Dentre os metadados de uma imagem, destacam-se suas coordenadas geográficas e suas *tags*, além de uma descrição e da identificação do proprietário da foto. Cada imagem contém zero ou mais *tags*, que são termos informados pelo usuário e que, para ele, descrevem a imagem ou momento em que a foto foi tirada. A Figura 2 apresenta um exemplo de fotografia e alguns de seus metadados. Outros metadados podem incluir a data de *upload* da foto no repositório, quantidade de visualizações e seu título.

2.1.2 Agrupamento de dados

O agrupamento de dados consiste em subdividir um conjunto de pontos, itens ou objetos em subconjuntos, de modo que cada subconjunto, chamado de grupo ou *cluster*, possua itens que sejam tão similares (ou próximos) entre si quanto possível e tão dissimilares (ou distantes) quanto possível daqueles pertencentes a outros grupos (HAN; KAMBER; PEI, 2011). Por não depender do conhecimento prévio dos grupos, sendo capaz de explorar

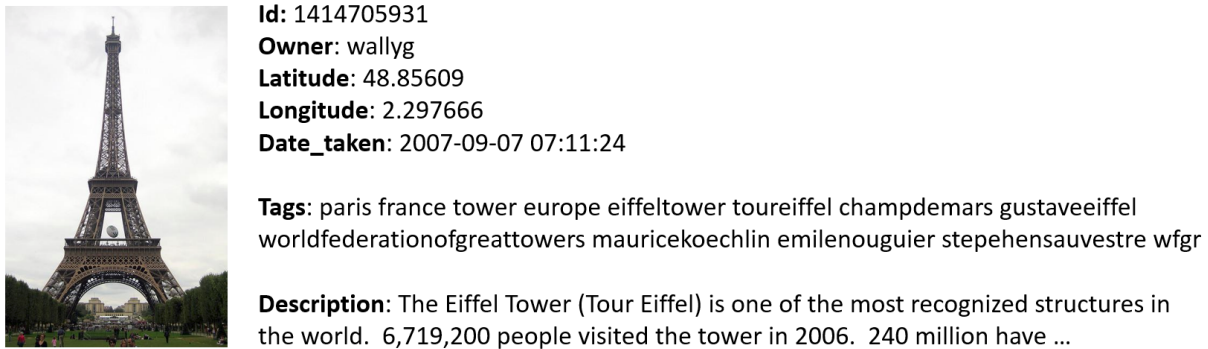


Figura 2 – Fotografia da Torre Eiffel (Paris, França) e alguns de seus metadados. Fonte: Flickr¹.

os dados e identificar os grupos automaticamente, o agrupamento é definido como uma tarefa não-supervisionada (JAIN, 2010).

Há vários algoritmos de agrupamento na literatura, os quais são divididos em categorias. Alguns métodos, inclusive, podem pertencer a mais de uma categoria de acordo com suas características. As principais, segundo (HAN; KAMBER; PEI, 2011), são apresentadas abaixo.

- ❑ **Métodos particionais:** Dado o número k de grupos a serem obtidos, os algoritmos constroem k grupos de dados iniciais. A partir daí, iterações são executadas de modo a mover objetos de um grupo para outro até que seja otimizado o critério de particionamento, que pode ser, por exemplo, uma função de dissimilaridade baseada na distância entre eles. Os algoritmos *k-Means* (MACQUEEN, 1967) e PAM (*Partitioning Around Medoids*) (KAUFMAN; ROUSSEUW, 1987) são exemplos de técnicas pertencentes a essa categoria.
- ❑ **Métodos hierárquicos:** Como o nome indica, métodos dessa categoria decompõem o conjunto de dados em uma estrutura hierárquica. Em sua abordagem aglomerativa (ou *bottom-up*), cada objeto do conjunto é considerado como um grupo. Assim, em cada iteração do método, os pares de grupos mais próximos são mesclados até que haja apenas um grupo (nível mais alto da hierarquia) ou uma condição de parada seja satisfeita. A abordagem divisiva (*top-down*), por outro lado, considera que todos os objetos formam um único grupo, que vai sendo dividido em grupos menores a cada iteração até que a estrutura de grupos desejada seja alcançada. São exemplos dessa categoria os métodos aglomerativos BIRCH (ZHANG; RAMAKRISHNAN; LIVNY, 1996), CURE (GUHA; RASTOGI; SHIM, 1998) e o algoritmo divisivo *Bisecting K-means* (STEINBACH; KARYPIS; KUMAR, 2000).

¹ <<https://www.flickr.com/photos/wallyg/1414705931/>> Acessado em 20/01/2016.

- ❑ Métodos baseados em grade: Dividem o espaço de dados em células, formando assim uma estrutura em forma de grade que independe da distribuição dos objetos que serão agrupados. Como o agrupamento geralmente depende apenas da quantidade de células e não da quantidade de objetos no conjunto de dados, a principal vantagem desse tipo de abordagem é o tempo de processamento. Os métodos STING (WANG; YANG; MUNTZ, 1997) e CLIQUE (AGRAWAL et al., 1998) são dois exemplos de técnicas dessa categoria.
- ❑ Métodos baseados em densidade: Consideram que grupos são regiões densas (alta concentração de objetos) separadas por regiões de baixa densidade (baixa concentração de objetos). A ideia básica consiste de, para cada objeto a ser agrupado, ir incluindo itens em seu grupo enquanto sua “vizinhança” for densa, ou seja, enquanto houver um número mínimo de objetos a uma dada distância máxima dele. Como apenas a distribuição dos objetos no espaço é considerada, a quantidade de grupos não é informada como entrada dos métodos e só é conhecida ao fim do processo de agrupamento. Métodos dessa categoria são capazes de encontrar grupos com formatos variados (Figura 3). Graças a essa característica, a categoria de métodos baseados em densidade recebe especial atenção no decorrer desta dissertação.



Figura 3 – Grupos com formatos variados. Imagem adaptada de (HAN; KAMBER; PEI, 2011).

A seguir, uma breve introdução sobre medidas de distância é apresentada e, na sequência, são descritos alguns métodos baseados em densidade.

2.1.2.1 Medidas de distância

Os métodos de agrupamento baseiam-se no valor de distância entre dois objetos para decidir se eles devem pertencer ao mesmo grupo. Os algoritmos de agrupamento por densidade, por exemplo, assumem um valor de distância máxima na análise da vizinhança de um dado objeto. Por essa razão, faz-se necessária a escolha de uma medida a ser utilizada no cálculo da distância entre os objetos.

Várias medidas de distância podem ser adotadas para esse fim. Uma das medidas mais conhecidas é a distância Euclidiana. Considerando os objetos $P = (p_1, p_2, \dots, p_n)$ e $Q = (q_1, q_2, \dots, q_n)$, a distância Euclidiana entre P e Q é dada por:

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

A distância Euclidiana é adequada quando os objetos sob análise estão em uma superfície plana (DEZA; DEZA, 2009). Nesse contexto, outras medidas também podem ser utilizadas, como a *City Block* ou a Chebyshev (WEBB, 1999). No entanto, em contextos geográficos, situações nas quais os objetos estão dispostos em uma superfície esférica (planeta Terra) e são representados pelas coordenadas geográficas latitude e longitude, a utilização dessas distâncias pode gerar erros, especialmente considerando-se objetos bem distantes entre si (IVIS, 2006). Isso ocorre pois essas medidas não levam em consideração a curvatura existente no espaço de dados.

Segundo (HIJMANS, 2015), existem quatro formas de se calcular a distância entre dois objetos geográficos, as quais são apresentadas em ordem crescente de complexidade algorítmica: lei esférica dos cossenos, Haversine (SINNOTT, 1984), “Vincenty *Sphere*” e “Vincenty *Ellipsoid*” (VINCENY, 1975). Do ponto de vista prático, os cálculos de distância efetuados com os três primeiros métodos geram resultados idênticos, sendo válido destacar que eles consideram a Terra como sendo uma esfera perfeita (HIJMANS, 2015).

Considerando os dois métodos mais simples, a fórmula de Haversine é matematicamente equivalente à lei esférica dos cossenos, sendo preferida por ser menos sensível a erros de arredondamento (IVIS, 2006). A distância de Haversine entre dois objetos $P = (lat_1, lng_1)$ e $Q = (lat_2, lng_2)$ é definida como se segue:

$$HaversineDist(P, Q) = R \cdot c$$

Onde R é o raio equatorial da Terra,

$$\begin{aligned} c &= 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{1-a}) \\ a &= (\sin(dlat/2))^2 + \cos(lat_1) \cdot \cos(lat_2) \cdot (\sin(dlng/2))^2 \\ dlat &= lat_2 - lat_1 \\ dlng &= lng_2 - lng_1 \end{aligned}$$

O resultado da equação estará na mesma unidade de medida que o valor de R , ou seja, se o raio da Terra for expresso em quilômetros, a distância entre P e Q obtida pela fórmula de Haversine também será expressa em quilômetros.

2.1.2.2 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

O DBSCAN (ESTER et al., 1996) é um dos métodos baseados em densidade mais populares. Sua análise exige o entendimento prévio de alguns conceitos. São eles:

- **Vizinhança de um ponto:** A vizinhança de um ponto p é composta por todos os pontos pertencentes à região delimitada pelo círculo de centro p e raio Eps (distância máxima aceitável para que dois pontos pertençam ao mesmo grupo). Cabe destacar que o próprio ponto é considerado como membro de sua vizinhança.
- **Ponto de núcleo (*core point*):** Um ponto é considerado de núcleo se o número de pontos em sua vizinhança Eps for superior a um limiar $MinPts$. Os valores de Eps e $MinPts$ são os dois parâmetros de entrada exigidos pelo método.
- **Ponto de borda (*border point*):** Um ponto de borda é um ponto que, apesar de não ser de núcleo, encontra-se na vizinhança de um. Um mesmo ponto de borda pode ter mais de um ponto de núcleo como vizinho.
- **Ruído (*noise point*):** É qualquer ponto que não seja de núcleo ou de borda.

O Algoritmo 1 apresenta os passos do DBSCAN (HAN; KAMBER; PEI, 2011).

Algoritmo 1: DBSCAN

Entrada: Conjunto de pontos; Eps ; $MinPts$

Saída: Conjunto de grupos

início

repita

 Seleciona-se aleatoriamente um ponto p não-visitado;

 Marca-se p como visitado;

se p é um ponto de núcleo **então**

 Cria-se um novo grupo C e insere-se p nele;

 Assume-se N como o conjunto de vizinhos de p ;

para cada ponto $p' \in N$ **faça**

se p' ainda não foi visitado **então**

 Marca-se p' como visitado;

se p' é um ponto de núcleo **então**

 Adiciona-se sua vizinhança em N ;

fim

se p' ainda não tem grupo **então**

 Adiciona-se p' em C ;

fim

fim

fim

 O grupo C está formado, insira-o na saída;

fim

até que todos os pontos tenham sido visitados;

fim

O DBSCAN é sensível aos parâmetros de entrada (Eps e $MinPts$), de forma que cada combinação entre eles afeta significativamente o resultado do agrupamento. Se, por

exemplo, ambos os parâmetros possuírem valores altos, o algoritmo resultará em poucos grupos grandes e densos. Se, por outro lado, ambos possuírem valores baixos, muitos grupos pequenos e pouco densos serão retornados. A escolha desses valores geralmente é feita de modo empírico (HAN; KAMBER; PEI, 2011).

Como mostra a Figura 4, a definição de valores únicos e globais de densidade *Eps* e *MinPts*, utilizados no DBSCAN, pode não ser adequada para representar corretamente conjuntos de dados com densidades variadas, situação vista em muitos dados reais (HAN; KAMBER; PEI, 2011). Essa limitação serviu como motivação para o desenvolvimento do OPTICS, método apresentado a seguir.

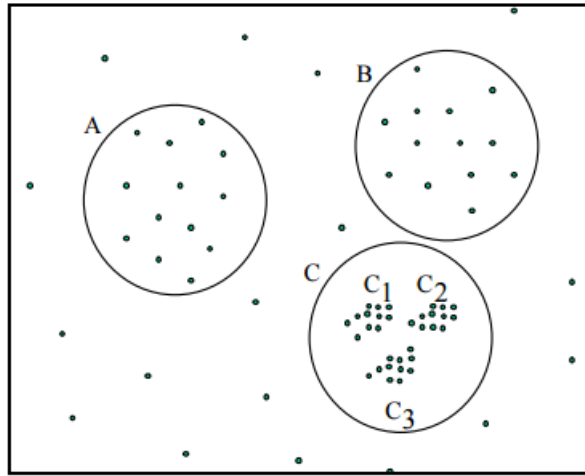


Figura 4 – Exemplo de grupos com diferentes densidades. A utilização de valores únicos e globais de densidade resultaria na obtenção dos grupos A, B e C. No entanto, o grupo C possui subgrupos que não seriam identificados (C1, C2 e C3). Imagem adaptada de (ANKERST et al., 1999).

2.1.2.3 OPTICS (*Ordering Points To Identify the Clustering Structure*)

Para contornar a dificuldade imposta pelo uso de um único conjunto de parâmetros globais no DBSCAN, os autores de (ANKERST et al., 1999) propuseram o OPTICS, método que funciona para infinitos valores de distância Eps_1 menores ou iguais ao Eps de entrada ($0 \leq Eps_1 \leq Eps$). Dessa forma, o algoritmo torna-se apto a identificar regiões com diferentes densidades. O OPTICS recebe como entrada os mesmos parâmetros do DBSCAN e utiliza o mesmo conceito de ponto de núcleo que ele.

O OPTICS não é explicitamente um método de agrupamento. Ao contrário do DBSCAN, sua saída é uma ordenação do conjunto de dados e não os grupos em si. Sua execução parte de duas definições (HAN; KAMBER; PEI, 2011):

- Distância de núcleo (*core distance*) de um ponto p : É a distância de p ao ponto vizinho que o torna um ponto de núcleo. Se p não é um ponto de núcleo, sua

distância de núcleo é considerada indefinida.

Exemplo de cálculo: assumindo-se $MinPts = 5$ e p um ponto de núcleo, tem-se que p tem pelo menos quatro vizinhos (já que um ponto é membro de sua vizinhança). Assim sendo, a distância de núcleo de p será a distância entre ele e seu quarto vizinho mais próximo.

- Distância de alcance (*reachability distance*) entre dois pontos p e q : Se p não é um ponto de núcleo, a distância de alcance é considerada indefinida. Caso contrário, a distância de alcance será o maior valor entre a distância de núcleo de p e a distância entre p e q . Nada impede que q possua mais de uma distância de alcance, cada uma relacionada a um ponto de núcleo.

O algoritmo mantém uma lista (chamada de *OrderSeeds*), onde, para cada elemento do conjunto de dados, são armazenadas os valores de distância de núcleo e distância de alcance. A lista é ordenada de acordo com a distância de alcance dos objetos aos pontos de núcleo mais próximos a eles, ou seja, de acordo com as menores distâncias de alcance de cada um (ANKERST et al., 1999).

O Algoritmo 2 apresenta os passos do método (HAN; KAMBER; PEI, 2011).

Algoritmo 2: OPTICS

Entrada: Conjunto de pontos; Eps ; $MinPts$

Saída: Conjunto ordenado de pontos

início

Define-se uma lista *OrderSeeds*, vazia;

Seleciona-se aleatoriamente um ponto p ;

repita

Encontra-se a vizinhança de p segundo Eps ;

Determina-se a distância de núcleo de p ;

Define-se sua distância de alcance como indefinida;

Coloca-se p na saída;

se p é um ponto de núcleo **então**

para cada ponto q em sua vizinhança **faça**

 Atualiza-se a distância de alcance de p a q ;

 Insere-se q em *OrderSeeds* caso ele ainda não tenha sido processado;

fim

fim

se *OrderSeeds* está vazia **então**

 Seleciona-se aleatoriamente um novo ponto p do conjunto de dados;

senão

 Define-se p como sendo o próximo ponto em *OrderSeeds*;

fim

até que todos os pontos tenham sido processados;

fim

Por fim, os grupos são montados percorrendo-se a lista resultante. Uma técnica proposta em (ANKERST et al., 1999) para esse fim é a plotagem dos pontos em um *Reachability Plot*. Dessa forma, os objetos pertencentes a uma mesma “zona de vale” na distribuição formam um grupo, como mostra a Figura 5. O primeiro elemento de um grupo será o último da sequência de valores altos para a distância de alcance. Isso é devido ao fato de a distância de alcance baixa ocorrer quando o elemento em questão está próximo ao seu antecessor no gráfico. Como exemplo, a Figura 6 exibe um grupo formado pelos elementos do intervalo [3, 16].

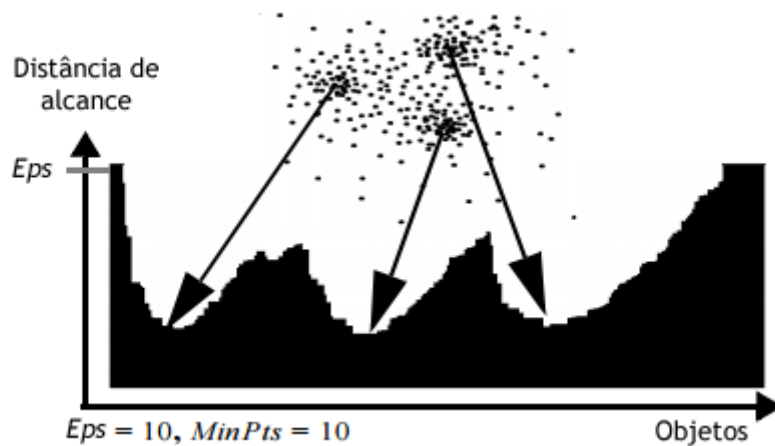


Figura 5 – Identificação dos grupos a partir da ordenação resultante. Imagem adaptada de (ANKERST et al., 1999).

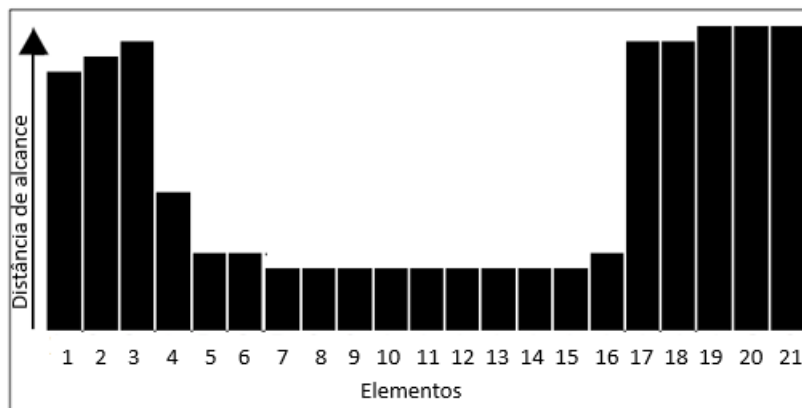


Figura 6 – Grupo composto pelos elementos do intervalo [3, 16]. Imagem adaptada de (ANKERST et al., 1999).

A estrutura do algoritmo OPTICS é bastante similar à do DBSCAN e, por essa razão, o tempo de execução de ambos é semelhante (HAN; KAMBER; PEI, 2011). A vantagem dele sobre o DBSCAN é que ele não se limita a um único valor de *Eps*, o que permite a identificação de grupos com diferentes densidades. Considerando um valor fixo para

$MinPts$, a estrutura ordenada obtida pelo OPTICS equivale a todas as execuções do DBSCAN para $0 \leq Eps_1 \leq Eps$.

2.1.2.4 Mean Shift

O *Mean Shift* (COMANICIU; MEER, 1999) é uma técnica de agrupamento simples e iterativa que identifica os grupos a partir do reconhecimento dos centros das regiões mais densas presentes no espaço de dados. Além dos dados que serão agrupados, o método recebe como entrada a distância máxima que define a vizinhança entre os objetos.

O trabalho (MONMA, 2014) resume o funcionamento básico do método da seguinte forma:

Algoritmo 3: *Mean Shift*

Entrada: Conjunto de pontos; d (distância máxima que considera dois objetos como vizinhos)

Saída: Conjunto de grupos

início

Escolhe-se um ponto aleatório p para compor o primeiro grupo;

repita

Define-se uma região de interesse como sendo o círculo de centro p e raio d ;

repita

Adiciona-se todos os pontos dessa região ao grupo;

Calcula-se a localização média dos pontos dessa região;

Desloca-se a região de interesse para a localização média, isto é, a localização média passa a ser o centro do círculo, dando origem a uma nova região de interesse;

até que a localização média não se altere significativamente (convergência do processo);

Escolhe-se um novo objeto aleatório p sem grupo;

Insere-se p em um novo grupo;

até que não haja ponto sem grupo;

fim

A execução do *Mean Shift* para um conjunto de dados sempre leva às regiões locais de máxima densidade (COMANICIU; MEER, 1999), conforme exemplifica a Figura 7.

Uma vantagem do *Mean Shift* é que, ao contrário do DBSCAN e OPTICS, apenas a distância máxima d (equivalente ao Eps nos dois métodos citados) é exigida como entrada do método. Assim como nos outros métodos, a escolha desse valor também influencia significativamente o resultado: um valor muito alto pode acabar mesclando grupos que deveriam ser distintos, ao passo que valores muito pequenos podem gerar grupos demais. A principal desvantagem do *Mean Shift* está em seu tempo computacional, que tende a

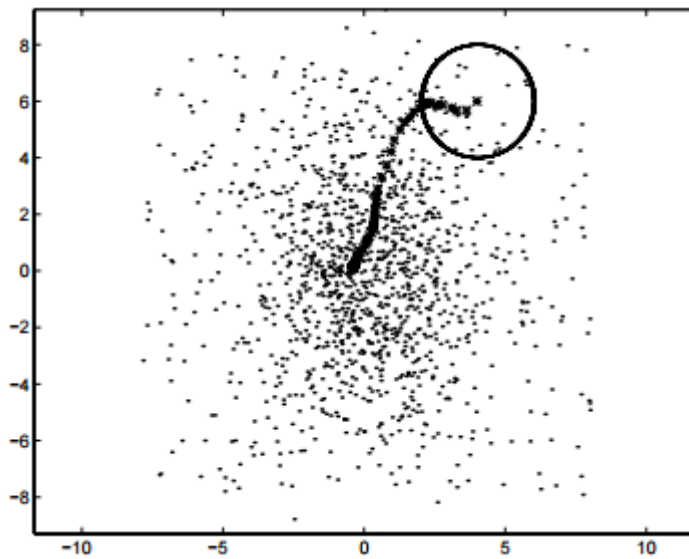


Figura 7 – Sucessivos deslocamentos de média definem um caminho que leva à máxima densidade local. Imagem adaptada de (COMANICIU; MEER, 1999).

ser superior ao dos demais algoritmos apresentados.

2.1.2.5 HDBSCAN (*Hierarchical DBSCAN*)

O HDBSCAN (CAMPELLO; MOULAVI; SANDER, 2013) é um método considerado por seus autores como uma melhoria do OPTICS. Ele recebe como parâmetro apenas a quantidade mínima de pontos por grupo (*MinPts*) e é capaz de identificar grupos em diferentes níveis de densidade (diferentes valores de *Eps*). Esse método utiliza o mesmo conceito de ponto de núcleo do DBSCAN e distância de núcleo do OPTICS. No entanto, ao contrário do DBSCAN, todos os pontos que não são de núcleo são considerados como ruído.

O HDBSCAN utiliza a definição de distância de alcance mútuo (*mutual reachability distance*), que é definida como $d_{AlcMutuo} = \max\{d_{nucleo}(p), d_{nucleo}(q), d(p, q)\}$, onde $d(p, q)$ é a distância entre p e q segundo a medida de distância sendo utilizada. A ideia desse cálculo é transformar o espaço, mantendo os pontos mais densos (com menores valores de distância de núcleo) com as mesmas distâncias entre si e afastando pontos esparsos. O Algoritmo 4 apresenta os passos do HDBSCAN (CAMPELLO; MOULAVI; SANDER, 2013; CAMPELLO et al., 2015).

Uma grande vantagem do HDBSCAN é a sua capacidade de reconhecer grupos em diferentes densidades, o que não ocorre com o DBSCAN. Assim como o *Mean Shift*, o HDBSCAN também exige apenas um parâmetro. Além disso, a definição do tamanho mínimo de um grupo (valor de *MinPts* exigido pelo HDBSCAN) tende a ser uma informação cuja determinação é mais intuitiva para o usuário do que a escolha da distância *Eps* exigida no *Mean Shift*.

Algoritmo 4: HDBSCAN**Entrada:** Conjunto de pontos; *MinPts***Saída:** Conjunto de grupos**início**

Calcula-se a distância de núcleo de todos os pontos;

 Calcula-se a distância de alcance mútuo entre todos os pares de pontos p e q ;

Representa-se o espaço na forma de um grafo onde os vértices são os pontos e as distâncias de alcance mútuo são os pesos das arestas entre os nós;

Obtém-se a árvore MST (GOWER; ROSS, 1969) desse grafo para se conhecer a estrutura hierárquica dos dados, que é dada pelo dendograma que corresponde à árvore;

Rotula-se todos os pontos da árvore como pertencentes ao mesmo grupo;

Ordena-se as arestas da árvore em ordem decrescente de peso;

para cada valor de peso ordenado **faça** Define-se o valor de escala do nível hierárquico corrente do dendograma como sendo o valor do peso corrente (esse valor corresponde a um *Eps*);

Remove-se a(s) aresta(s) com o valor de peso corrente;

para cada grupo que contém a(s) aresta(s) removida(s) **faça** **se** um subcomponente dele possuir menos pontos que o valor de *MinPts* **então**

| Marca-se os “filhos” como ruído, descartando-os;

senão se um (e apenas um) subcomponente possui *MinPts* pontos ou mais **então**

| Mantém-se o subcomponente como pertencente ao grupo original;

senão se se dois ou mais subcomponentes possuem *MinPts* pontos ou mais **então**

| Associa-se cada subcomponente a um novo grupo;

fim **fim** **fim****fim****2.1.2.6 P-DBSCAN (“Photo” DBSCAN)**

Ao contrário dos métodos baseados em densidade apresentados, cujas características permitem o agrupamento de dados d-dimensionais provenientes de diversos contextos, o P-DBSCAN (KISILEVICH; MANSMANN; KEIM, 2010) é um método de agrupamento derivado do DBSCAN que tem como objetivo o agrupamento de fotografias a partir de informações de geo-localização e de usuários a fim de identificar locais de interesse. Sua execução faz uso de algumas definições:

1. Vizinhança de uma foto: a vizinhança de uma foto p é formada por fotos tiradas a uma distância máxima *Eps* e por usuários diferentes do proprietário de p .
2. Foto de núcleo: uma foto é considerada de núcleo se existem ao menos *MinOwners* usuários em sua vizinhança;

3. Limiar de densidade adaptativa (*adaptive density threshold* - *Addt*): Proporção entre a densidade atual e a anterior. Os vizinhos de uma foto vão sendo incluídos em um mesmo grupo até que a densidade alcançada seja maior que esse limiar. A entrada desse limiar no algoritmo é opcional e determinada por um valor lógico *Ad* (o limiar é utilizado se *Ad* é verdadeiro).

O Algoritmo 5 descreve os passos do P-DBSCAN.

Algoritmo 5: P-DBSCAN

Entrada: Conjunto de fotos e seus atributos; *Eps*; *Ad*; *Addt*

Saída: Conjunto de grupos

início

repita

 Escolhe-se uma foto *p* não processada;

 densidadeCorrente = *Addt*;

se a quantidade de vizinhos de *p* for menor que *MinOwners* **então**

 | Marca-se *p* como ruído;

senão

 Cria-se um novo grupo *C* e insere-se *p* nele;

 Assume-se *N* como o conjunto de vizinhos de *p*;

para cada ponto *p'* ∈ *N* **faça**

 | Insere-se *p'* em *C*;

se a quantidade de vizinhos de *p'* for maior ou igual que *MinOwners*

então

 | **se** *Ad* é falso **então**

 | | Insere-se os vizinhos de *p'* em *N*;

 | **senão**

 | | limiarDeDensidade = (quantidade de vizinhos de *p'*) /
 | | densidadeCorrente;

 | **se** a quantidade de vizinhos de *p'* for menor que
 | | densidadeCorrente **então**

 | | | **se** limiarDeDensidade ≥ *Addt* **então**

 | | | | densidadeCorrente = quantidade de vizinhos de *p'*;

 | | | | Insere-se os vizinhos de *p'* em *N*;

 | | **fim**

 | **senão**

 | | densidadeCorrente = quantidade de vizinhos de *p'*;

 | | Insere-se os vizinhos de *p'* em *N*;

 | **fim**

fim

fim

fim

fim

até que todas as fotos tenham sido processadas;

fim

2.2 Trabalhos Correlatos

Vários trabalhos na literatura utilizam dados geo-referenciados provenientes de redes sociais para identificar pontos de interesse turísticos. Cao et al. (CAO et al., 2010) propuseram um sistema de recomendação turística que utiliza cerca de um milhão de fotos, obtidas no Flickr², que possuem geo-localização e outros metadados. Nesse sistema, PoIs são identificados com o uso do algoritmo *Mean Shift* para agrupar as imagens de acordo com suas coordenadas GPS. Cada grupo obtido é considerado um local de interesse e é representado por um conjunto com as *tags* mais frequentes nele e pelo centro do grupo em termos de latitude e longitude.

Em (CRANDALL et al., 2009), os autores executaram o *Mean Shift* em cerca de 34 milhões de imagens extraídas do Flickr com o objetivo de identificar PoIs em uma escala global. Para obter as *tags* mais discriminantes de cada PoI, os autores descartaram todas as *tags* que não tivessem ocorrido em, pelo menos, 5% das fotos do respectivo grupo. Posteriormente, para cada *tag* restante do grupo, calculou-se a taxa entre o número de fotos no grupo com essa *tag* e a quantidade total de fotos que possuem tal *tag*. Ao fim do processo, a *tag* com a maior taxa foi escolhida para representar o local.

Os autores de (AHERN et al., 2007) apresentaram o sistema *World Explorer*, uma ferramenta de visualização que mostra *tags* representativas em qualquer região geográfica do mundo (Figura 8). Para isso, os autores propuseram agrupar as fotografias com o método particional *k-Means*, aplicando-o às coordenadas geográficas das imagens. Para cada grupo resultante, todas as *tags* foram associadas a uma pontuação calculada com o TF-IDF-UF, um algoritmo criado pelos autores baseando-se na clássica versão da medida estatística *Term Frequency-Inverse Document Frequency* (TF-IDF) (SALTON; BUCKLEY, 1988) e que leva em consideração a porcentagem de usuários no grupo que utilizaram a *tag* em análise. Posteriormente, os autores do trabalho (HE et al., 2012) estenderam o TF-IDF-UF para também considerar a distribuição geográfica da *tag*, o que resultou em uma abordagem de pontuação com melhores resultados.

Os trabalhos mencionados fazem uso apenas das coordenadas GPS e de uma ou mais *tags* para representar cada local identificado. Por outro lado, algumas propostas utilizam a Wikipedia³ e, assim, possuem mais informações sobre os PoIs, como nome do local, categoria e descrição (BRILHANTE et al., 2013; SHI et al., 2013). Em (BRILHANTE et al., 2013), por exemplo, cada página da Wikipedia com geo-localização situada na cidade escolhida para o reconhecimento de PoIs é considerada uma entidade. O método de agrupamento DBSCAN é então utilizado para agrupar entidades próximas de modo que cada grupo resultante seja considerado um PoI. Fotos da cidade em questão são recuperadas do Flickr e cada foto é associada a um PoI se sua distância a ele for de, no máximo, 100 metros. Ressalta-se que as cidades passíveis de escolha pelo usuário são

² <https://www.flickr.com>

³ www.wikipedia.org

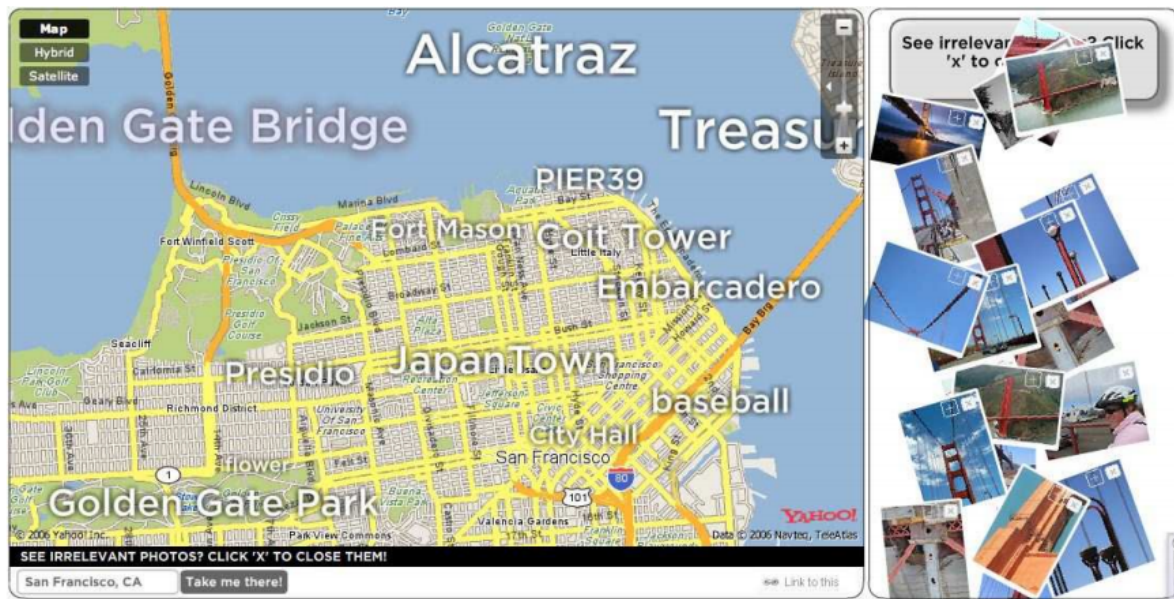


Figura 8 – Tela do sistema *World Explorer* exibindo partes de São Francisco, EUA. O usuário selecionou a tag “Golden Gate Bridge” (porção superior à esquerda) para que fossem exibidas imagens com essa tag nessa região geográfica. Imagem adaptada de (AHERN et al., 2007).

pré-determinadas em Roma, Florença, Pisa e Amsterdã (BRILHANTE et al., 2014b), fato que restringe o uso da aplicação. Além disso, lugares interessantes podem não ser reconhecidos por essa estratégia se não possuírem nomes bem definidos, como é o caso de bairros com arte de rua, pois, nesse caso, poderia não haver páginas na Wikipedia a respeito deles. O usuário pode optar por visitar tais lugares e, por essa razão, considerar que PoIs são entidades da Wikipedia pode não ser sempre adequado para a identificação de pontos turísticos.

No sistema de recomendação proposto pelos autores de (MAJID et al., 2015), PoIs são automaticamente identificados por meio do agrupamento das fotografias aplicando-se o método P-DBSCAN às coordenadas geográficas das imagens. Para cada grupo, todas as tags são ordenadas em ordem decrescente de acordo com a pontuação TF-IDF-UF proposta em (AHERN et al., 2007). Cada grupo é considerado um PoI e informações sobre ele são recuperadas consultando-se o serviço de busca por proximidade geográfica Google Places⁴. Se houver, dentre os locais retornados pelo serviço, um local cujo nome combine com uma das tags do grupo, seu nome é associado ao PoI; se houver mais de um nome que combine, o PoI é associado ao local que for mais próximo geograficamente a ele; por fim, se não houver nenhuma combinação, o nome do PoI é associado à tag com maior pontuação.

Os trabalhos citados acima identificam PoIs utilizando estratégias diferentes. No en-

⁴ www.google.com/places

tanto, nenhum deles aborda essa questão em um contexto dinâmico, isto é, essas abordagens podem, potencialmente, exibir para o usuário um PoI que não existe mais ou não reconhecer novos PoIs. A percepção de aparecimento ou desaparecimento de PoIs é comumente observada em mídias sociais, como o Flickr, onde várias novas fotos são postadas ao longo do tempo. Se a frequência de fotos tiradas em um determinado local decresce com o tempo, ou mesmo desaparece, pode significar que aquele PoI não é mais interessante por diferentes razões ou sequer existe mais. Da mesma forma, se a frequência de fotos cresce ao longo do tempo em um local onde não havia nada interessante antes, um novo PoI pode estar presente ali.

Uma abordagem que tira proveito da dinamicidade do Flickr foi proposta em (GAGGI, 2013), mas o trabalho não faz distinção quanto às datas em que as fotos recuperadas foram tiradas, o que dificultaria, por exemplo, a identificação de novos PoIs. Apesar de oferecer uma busca por locais a uma distância máxima de determinado endereço (tal qual a proposta deste trabalho de mestrado), o objetivo do autor é retornar PoIs pouco conhecidos, especialmente aqueles próximos à rota de viagem entre dois endereços informados pelo usuário. Uma limitação desse trabalho é que, para retornar apenas locais pertencentes à categoria escolhida pelo usuário (como arquitetura, por exemplo), a abordagem requer que as categorias de PoIs sejam explicitamente representadas por *tags*, informação essa que nem sempre está disponível. Além disso, essa associação é fortemente dependente do idioma utilizado, o que restringe a utilização da proposta.

2.3 Considerações Finais

Este capítulo apresentou conceitos relevantes para o trabalho e descreveu, sem detalhes, o funcionamento de alguns métodos de agrupamento de dados que podem ser usados, no contexto deste trabalho, para a identificação de pontos turísticos a partir de imagens com metadados. A revisão da literatura apresentou diversas abordagens e expôs o problema da identificação de PoIs em contexto dinâmico, representado pela dinamicidade de repositórios *online* de imagens e que não é tratado nos trabalhos citados. Os capítulos seguintes apresentam a proposta deste trabalho, desenvolvido para aplicação nesse cenário.

Método ToPI de identificação de locais de interesse

A identificação de pontos de interesse é uma etapa comum a vários trabalhos que atuam na área de turismo. Apesar de muitas abordagens utilizarem fotografias oriundas de repositórios *online* (como o Flickr), as estratégias apresentadas no capítulo anterior não consideram a dinamicidade inerente a esse tipo de fonte de dados, na qual fotografias são publicadas e retiradas o tempo todo, na identificação de pontos turísticos populares.

Este capítulo apresenta o método *Tourist Place Identification* (ToPI), desenvolvido para identificar e rotular PoIs a partir de informações obtidas em repositórios *online* de fotografias, mais precisamente a partir da recuperação em tempo real de imagens com geo-localização e outros metadados associados. A dinamicidade inerente ao problema é naturalmente tratada pelo método proposto, que, além de identificar PoIs permanentes, também é capaz de identificar lugares turísticos sazonais.

3.1 ToPI

A Figura 9 apresenta a visão geral do processo, mostrando as etapas necessárias para identificar PoIs na região geográfica desejada. O processo pode ser sumarizado em módulos como se segue:

1. Seleciona-se, em um repositório *online* de fotografias, fotos que pertençam a uma dada região geográfica e seus metadados;
2. Agrupa-se as imagens obtidas de acordo com suas coordenadas GPS (latitude e longitude) e filtra-se os grupos resultantes de modo a descartar aqueles com poucos usuários - a partir deste ponto, cada grupo é assumido como sendo um PoI;
3. Identifica-se palavras-chave que melhor descrevem cada PoI;

4. Trata-se as redundâncias de PoIs, de modo que cada local de interesse seja representado por apenas uma ocorrência;
5. Obtém-se detalhes a respeito de cada PoI em uma fonte de informações externa;
6. Armazena-se os PoIs e suas informações em um banco de dados.

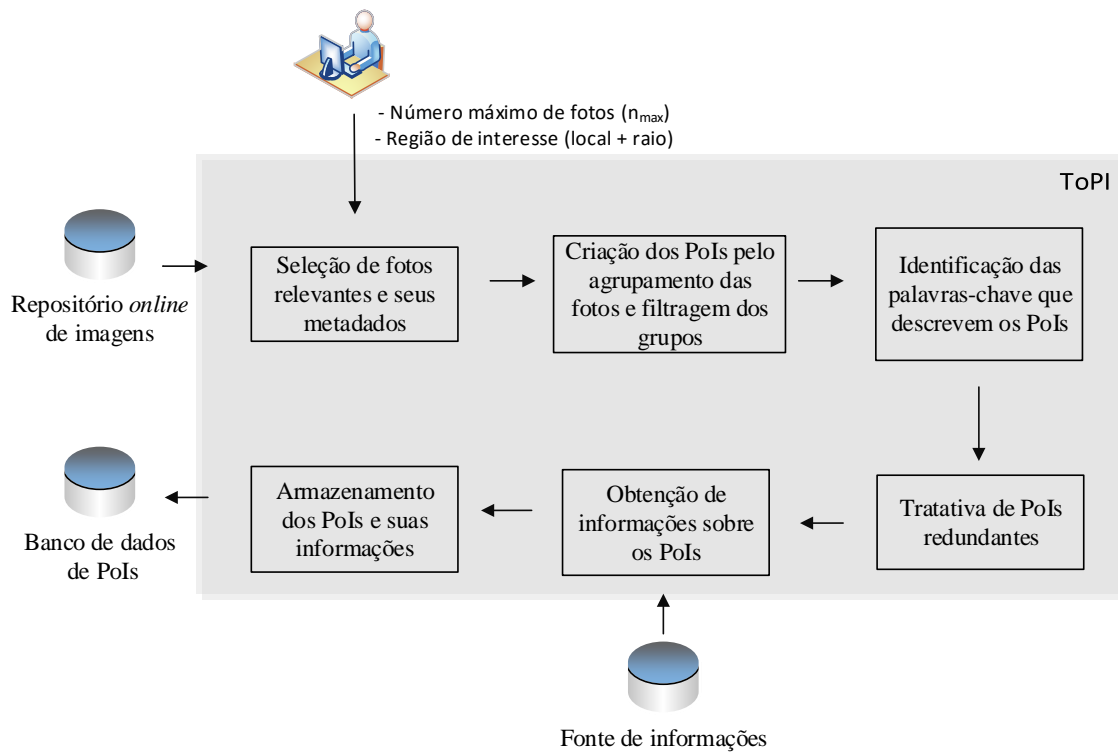


Figura 9 – ToPI - Uma visão geral do fluxo de dados.

Cada módulo do método proposto será apresentado em detalhes a seguir.

3.1.1 Seleção das fotografias

Este módulo tem como objetivo selecionar até n_{max} fotos geo-referenciadas e seus respectivos metadados de um repositório *online*. Essas fotos serão utilizadas na identificação de pontos de interesse, dada uma localização em termos de coordenadas geográficas e um raio de alcance.

Apesar de não ser obrigatória a utilização de fotografias recentes para que seja feita a identificação de PoIs, a recuperação e uso desse conjunto de imagens possibilita a percepção de situações como o fechamento e abertura de locais de interesse, além da percepção de sazonalidades.

Para cada imagem selecionada, os metadados armazenados são: lista de *tags* (contendo zero ou mais *tags*), a identificação da foto no repositório *online*, coordenadas geográficas em termos de latitude e longitude, data e hora em que a foto foi tirada, nome de usuário do proprietário da foto e a quantidade de visualizações que ela possui.

3.1.2 Agrupamento e filtragem

Neste módulo, as fotografias são inicialmente agrupadas conforme suas coordenadas geográficas. A concentração de fotos tiradas em um mesmo local indica que ele é um lugar potencialmente turístico. Essa estratégia foi inspirada no trabalho (HE et al., 2012) e permite assumir um PoI como um grupo de fotografias.

A Figura 10 mostra um conjunto de 500 imagens selecionadas nas proximidades do Cristo Redentor (Rio de Janeiro, RJ, Brasil). Qualquer técnica de agrupamento pode ser utilizada nesta etapa. No entanto, como não é possível estimar previamente quantos PoIs existem em uma dada área geográfica, um método de agrupamento baseado em densidade pode representar uma opção mais eficaz para identificar as aglomerações de fotos a partir de suas coordenadas GPS.

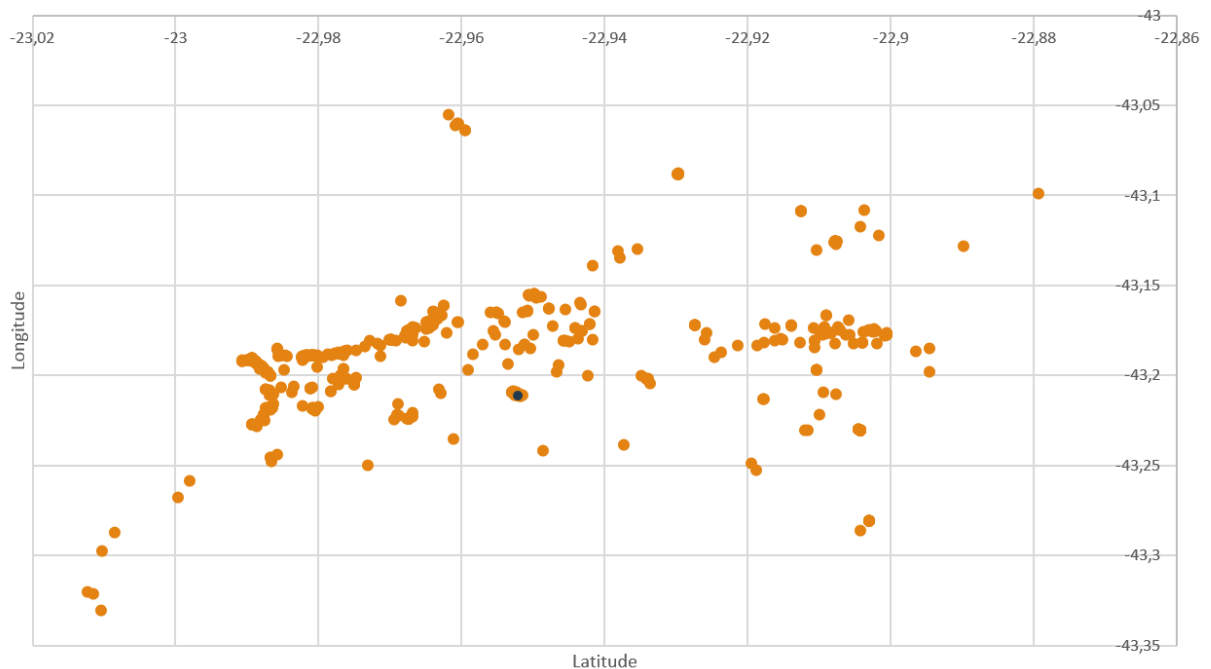


Figura 10 – Gráfico de dispersão com 500 fotografias recuperadas nas proximidades do Cristo Redentor (representado pelo ponto destacado em azul no centro da imagem).

Uma vez obtido o agrupamento, uma filtragem é realizada para eliminar grupos que possuem menos que T_{users} usuários. Nesse caso, o grupo é entendido como ruído e não é considerado como um PoI. Um exemplo dessa situação é a presença de um grupo composto

por fotos de um jardim tiradas por seu proprietário. Após essa etapa de filtragem, cada grupo resultante é considerado um ponto de interesse e pode ter sua localização calculada. Daqui em diante, os termos “grupo” e “PoI” serão usados indistintamente ao longo do trabalho.

A localização geográfica de cada PoI, dada pela posição de seu centro representativo (P_{CR}), é calculada da seguinte forma: inicialmente, calcula-se a localização média das fotos de cada usuário do grupo; em seguida, realiza-se uma média global a partir da média de cada usuário. Esse processo é realizado dessa maneira para evitar que um usuário com muitas fotografias em um grupo tenha maior influência na determinação do P_{CR} .

3.1.3 Identificação das palavras-chave que descrevem os PoIs

Os grupos resultantes do passo anterior representam locais de interesse na região geográfica em análise. Entretanto, a única informação em mãos até o momento é a localização do centro representativo de cada PoI. Cada grupo é composto por fotos textualmente representadas por suas *tags*. Cada foto pode, no entanto, possuir *tags* que não ajudariam a reconhecer o local, como nomes de pessoas, modelo da câmera fotográfica utilizada, etc. Essas *tags* podem ser descartadas sem prejuízo para a representação da imagem e, conseqüentemente, do grupo. Assim, a fim de identificar aquelas *tags* que melhor discriminam um PoI, uma pontuação é calculada e associada a cada *tag* de cada imagem dentro do grupo. Um critério de pontuação possível é pontuar determinada *tag* com base na quantidade de usuários do grupo em questão que a utilizaram em suas imagens. Com isso, quanto maior a pontuação atribuída, mais distintiva a *tag* é no grupo.

As *tags* de um grupo são ordenadas decrescentemente segundo suas pontuações e as *tags* com maiores pontuações são usadas para determinar as palavras-chave que possam ajudar no processo de identificação do respectivo PoI. Este trabalho assume que existe uma relação entre o tamanho de uma *tag* (em termos de quantidade de caracteres) e sua capacidade de descrever o PoI. Assim, a maior *tag* que possui a *tag* de maior pontuação como *substring* é considerada como sendo a primeira palavra-chave. Entretanto, às vezes essa palavra-chave se torna muito específica para alcançar um correto reconhecimento do PoI e, por esta razão, a *tag* original é mantida como segunda palavra-chave. Essa estratégia é aplicada a todas as *tags* da lista ordenada previamente obtida ignorando-se as palavras-chave repetidas.

As Tabelas 1 e 2 apresentam exemplos de identificação de palavras-chave assumindo-se a existência de quatro *tags* no grupo em questão. A partir da Tabela 1, pode-se observar que a palavra-chave que melhor representa o local é a primeira (“torreeiffel”). Assim, a primeira tentativa de identificação do local teria mais chances de êxito. Na Tabela 2, pode-se perceber que a primeira palavra-chave é (provavelmente) específica demais, a segunda é (provavelmente) genérica demais e a terceira é a que melhor representa o local. Neste caso, a identificação do local possivelmente teria êxito apenas na terceira tentativa.

Tabela 1 – Identificação de palavras-chave. Exemplo 1.

Pontuação	Tags existentes	Palavras-chave
2,55	eiffel	torreeiffel
2,4	torre	eiffel
2	torreeiffel	torre
0,9	monument	monument

Tabela 2 – Identificação de palavras-chave. Exemplo 2.

Pontuação	Tags existentes	Palavras-chave
2,55	eiffel	atorreeiffelbonita2015
2,4	torre	eiffel
1	atorreeiffelbonita2015	torreeiffel
0,9	torreeiffel	torre

3.1.4 Tratativa de PoIs redundantes

A etapa de agrupamento pode resultar em mais de um grupo representando o mesmo local, ou seja, PoIs redundantes. Essa é uma característica recorrente em lugares com vários pontos de vista (onde cada ponto de vista é representado por um grupo distinto) e em locais bastante extensos, como praias e parques, nos quais é comum haver fotos em diferentes regiões. Essas redundâncias precisam ser removidas a fim de evitar que o usuário se confunda achando que trata-se de PoIs distintos.

Representando um PoI x por P^x , sua localização por P_{CR}^x e assumindo-se N_{pc}^x como sendo o conjunto das n primeiras palavras-chave (pc) de P^x , todos os PoIs que se encaixam em uma das duas regras a seguir são unidos.

1. $dist(P_{CR}^a, P_{CR}^b) \leq T_{MergeD}$, onde $dist(X, Y)$ é a função de distância sendo utilizada e T_{MergeD} é um limiar que indica a distância máxima para que dois PoIs sejam considerados muito próximos. A ideia desse critério é que PoIs em uma região muito pequena referem-se ao mesmo local.
2. $T_{MergeD} < dist(P_{CR}^a, P_{CR}^b) \leq T_{MergeK}$ e pelo menos uma palavra-chave pertencente a N_{pc}^a contém ou está contida em alguma palavra-chave de N_{pc}^b . A ideia é que grupos em uma região de tamanho médio referem-se ao mesmo PoI se pelo menos uma das palavras-chave mais relevantes forem coincidentes. Este critério parte do princípio de que as palavras-chave mais relevantes são as primeiras (representadas por N_{pc}), já que a ordem de obtenção delas segue a ordem decrescente de pontuação das *tags*.

A Figura 11(a) apresenta um exemplo de união de dois PoIs segundo a regra 1. Nesse caso, a união é feita independentemente das palavras-chave que representam os dois PoIs. Já a Figura 11(b) traz um exemplo de aplicação da regra 2. Nesse exemplo, os PoIs 3 e 4

só seriam unidos se possuísem pelo menos uma palavra-chave coincidente nos conjuntos N_{pc}^3 e N_{pc}^4 . Cabe destacar que o PoI 5 não seria unido a nenhum outro PoI, mesmo que houvesse palavra(s)-chave coincidente(s).

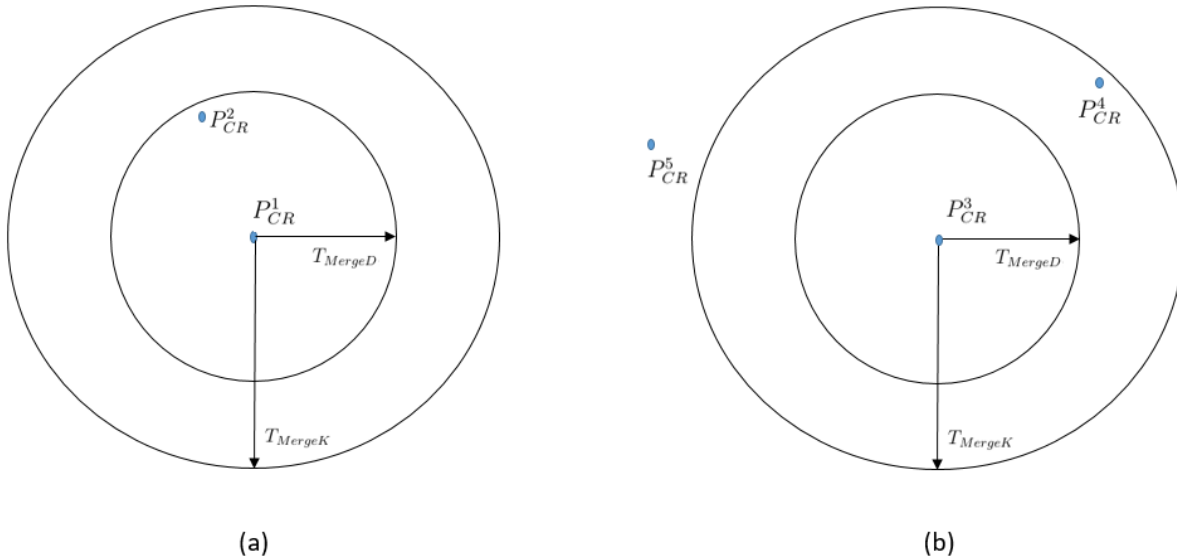


Figura 11 – Exemplo de união entre PoIs: (a) União automática entre os PoIs 1 e 2 (regra 1); (b) União entre os PoIs 3 e 4 apenas se existir palavra-chave coincidente nos conjuntos N_{pc}^3 e N_{pc}^4 (regra 2).

A análise de PoIs redundantes é realizada recursivamente e o processo de união analisa os PoIs dois a dois até que nenhuma das regras acima se aplique aos grupos resultantes. Cada iteração recalcula as pontuações das *tags* pertencentes ao novo grupo sendo formado, de modo que novas palavras-chave possam surgir e alimentar a iteração seguinte.

3.1.5 Obtenção de informações sobre os PoIs

Após o processo de remoção de PoIs redundantes, as primeiras palavras-chave de cada PoI (N_{pc}) são usadas para recuperar o nome e a categoria (museu, parque, etc.), além de outras informações potencialmente relevantes relacionadas. Essas palavras-chave são usadas sequencialmente em consultas a uma fonte de informações, que pode estar disponível *online* ou *offline*, até que um resultado válido seja obtido. A busca pode retornar dois ou mais PoIs com nomes similares e localizações diferentes, como mostrado na Figura 12. Nesse caso, o resultado é considerado válido se o nome do local retornado equivale à palavra-chave utilizada e se a distância entre o centro representativo do PoI e o local retornado for igual ou menor a um limiar $T_{maxdist}$.

Se as fotos de um grupo não possuem nenhuma *tag* ou todas as palavras-chave de N_{pc} foram usadas e nenhum resultado válido foi obtido, as coordenadas geográficas do PoI são

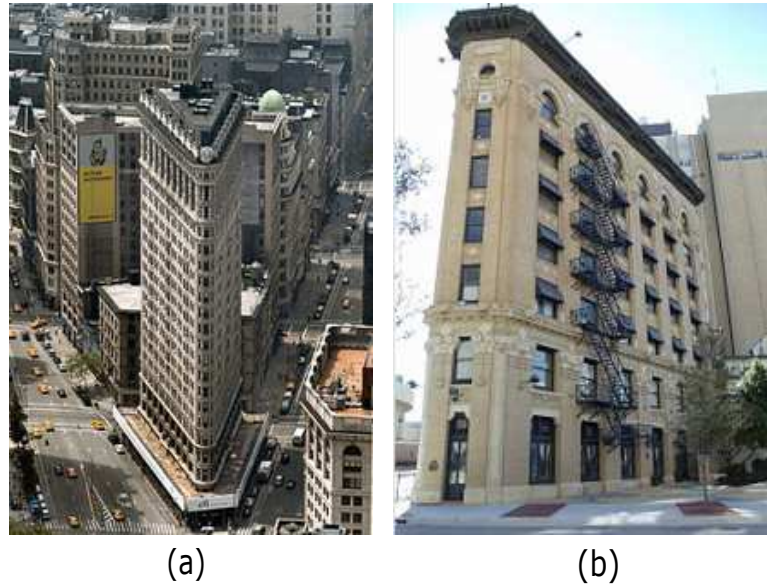


Figura 12 – Dois prédios chamados “Flatiron Building” em diferentes cidades dos Estados Unidos: (a) New York City, New York; (b) Fort Worth, Texas.

utilizadas em uma busca por proximidade geográfica em uma fonte de informações que ofereça esse tipo de pesquisa via uma consulta com raio igual a $T_{maxdist}$.

A busca por meio das palavras-chave recebe prioridade em relação à busca por proximidade geográfica devido à situação apresentada na Figura 13. Nesse exemplo, tem-se dois PoIs em Nova York: *Madison Square Park* e *National Museum of Mathematics*. Assumindo-se hipoteticamente que os círculos azuis representam as coordenadas geográficas deles na fonte de informações e que os círculos vermelhos representam seus centros representativos calculados pelo método (P_{CR}), a busca por proximidade geográfica rotularia o parque como sendo o museu, já que o P_{CR} do parque está mais próximo da localização conhecida do museu do que de sua própria localização. Assim, para evitar associações indevidas, especialmente quando o P_{CR} de um PoI estiver nas extremidades do local, a busca por proximidade só é utilizada como última tentativa na obtenção de informações.

Outra informação obtida para cada PoI é a sua popularidade relativa aos demais, que é calculada com base na quantidade de usuários distintos no grupo: se o PoI em questão possuir duas vezes ou mais a quantidade média de usuários por grupo identificado, sua popularidade será de 100%; caso contrário, sua popularidade assumirá um valor proporcional.

Ao fim do processo, cada PoI é associado a um valor de popularidade, nome, categoria e a um conjunto com as fotos mais visualizadas daquele grupo. Quando as buscas por palavras-chave e por proximidade falham, o PoI é associado apenas à popularidade e ao conjunto de fotos. Por fim, os PoIs e suas informações são armazenados em um banco de dados de PoIs para serem utilizados posteriormente.

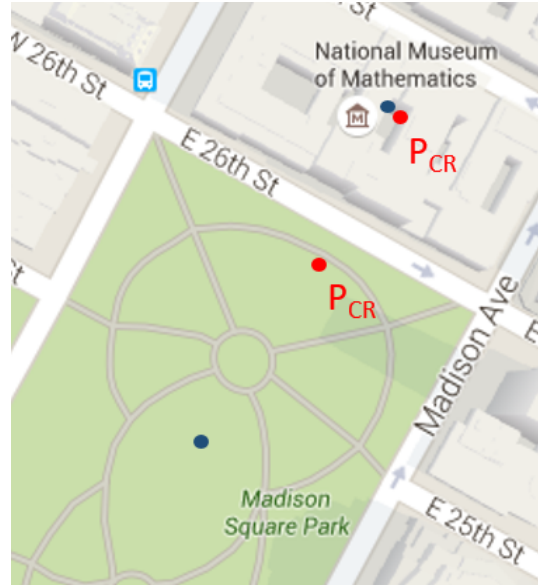


Figura 13 – Exemplo que destaca a importância de se priorizar a busca textual em detrimento da busca por proximidade geográfica na obtenção de informações de um PoI.

3.2 Considerações Finais

Este capítulo descreveu o método ToPI de identificação de pontos turísticos. O método é flexível e permite a utilização de qualquer repositório *online* de imagens que forneça geolocalização e os outros metadados. Além disso, pode ser utilizado qualquer método de agrupamento e qualquer fonte de informação que ofereça busca textual e/ou busca por proximidade geográfica.

O capítulo a seguir apresenta o sistema ToPI *Trip*, um sistema de identificação de PoIs, disponível via *Web*, que utiliza o método ToPI descrito neste capítulo.

Sistema ToPI Trip

Este capítulo descreve o sistema *ToPI Trip*, uma aplicação Web desenvolvida para avaliar o método proposto. A Seção 4.1 apresenta uma visão geral do sistema. A Seção 4.2 descreve, em detalhes, as funcionalidades disponíveis nele. A Seção 4.3 apresenta os fluxos executados de modo a obter uma experiência em tempo real. A Seção 4.4 discorre sobre a escolha do repositório de fotografias, o método de pontuação das *tags* escolhido e os critérios considerados na escolha do método de agrupamento utilizado nos experimentos do capítulo seguinte. Por fim, a seção 4.5 apresenta as considerações finais a respeito do capítulo.

4.1 Visão geral do sistema

A Figura 14 apresenta o sistema *ToPI Trip* sob três perspectivas: identificação de PoIs e informações, responsável pela execução do método ToPI e pela obtenção de outras informações relacionadas; visualização, responsável pela exibição do conhecimento obtido; e interação do usuário, com ferramentas que permitem a exploração desse conhecimento.

O sistema *ToPI Trip* oferece:

- ❑ Busca por PoIs em qualquer localidade;
- ❑ Exibição dos PoIs em um mapa interativo;
- ❑ Visualização de informações básicas dos PoIs, como nome, popularidade, categoria e fotografias do local;
- ❑ Visualização de estatísticas de visitação dos PoIs no ano anterior;
- ❑ Escolha de um valor de popularidade mínima para a exibição dos PoIs;
- ❑ Integração com o *Bing Knowledge Widget*¹, que permite buscar e visualizar mais informações sobre os PoIs e sobre a cidade e país da localização informada;

¹ www.bing.com/widget/knowledge

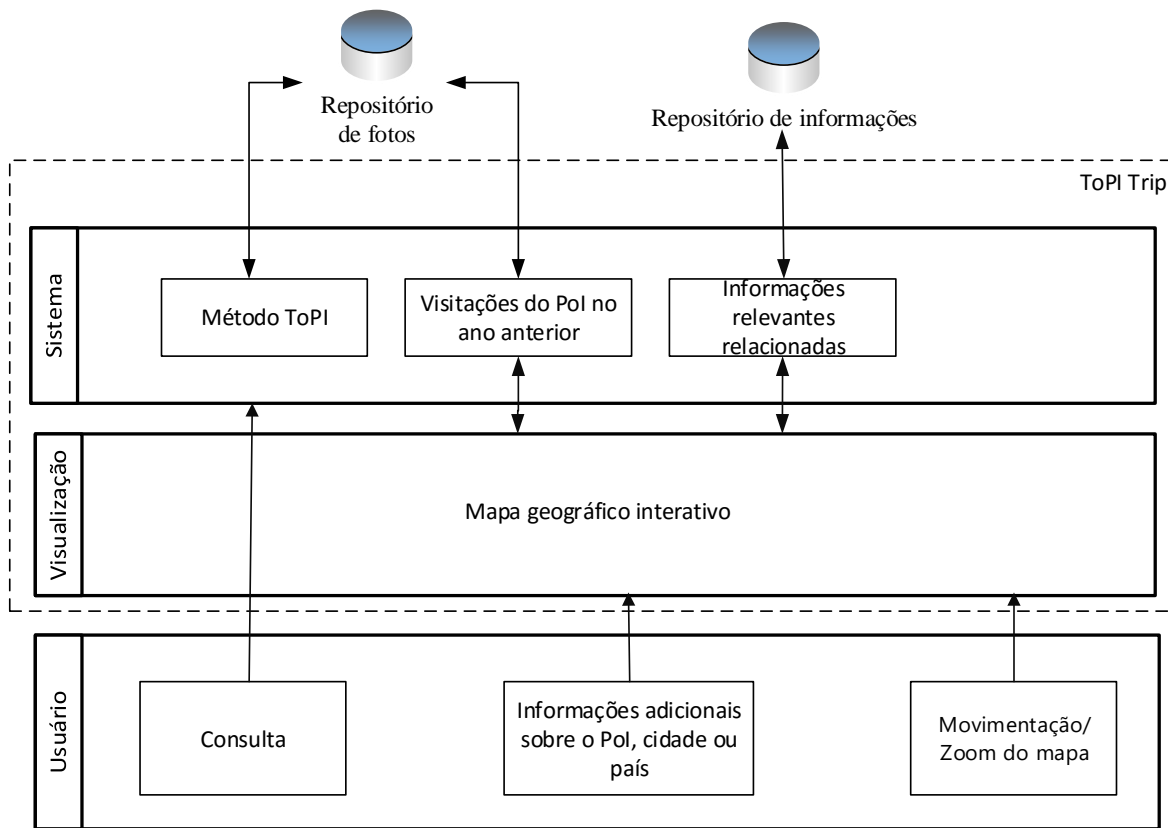


Figura 14 – Visão geral do sistema ToPI *Trip*.

- ❑ Visualização das distâncias entre os locais exibidos.

As funcionalidades disponíveis no sistema são apresentadas abaixo com mais detalhes.

4.2 Funcionalidades do sistema ToPI *Trip*

O sistema ToPI *Trip* é modularizado e pode ser estendido de forma a utilizar qualquer repositório que ofereça fotografias com geo-localização e demais metadados. O Flickr foi escolhido para este trabalho e, assim, sua Interface de Programação de Aplicativos - *Application Programming Interface* (API) (Flickr, 2015) foi utilizada na recuperação de imagens tiradas dentro da área geográfica definida pela localização informada pelo usuário e pelo raio de abrangência da busca. Os critérios que levaram à escolha do Flickr serão detalhados na Seção 4.4.1.

Utilizando as fotografias selecionadas no repositório *online* como entrada, o sistema ToPI *Trip* exibe todos os PoIs identificados em uma mapa com vários recursos de interação. A Figura 15 apresenta alguns PoIs retornados em uma busca por locais de interesse em Paris, França. O marcador com símbolo de uma bandeira, na porção esquerda da imagem,

no repositório. Assim, ele pode, por exemplo, salvar a imagem em seu computador ou explorar outras fotografias do proprietário da foto.

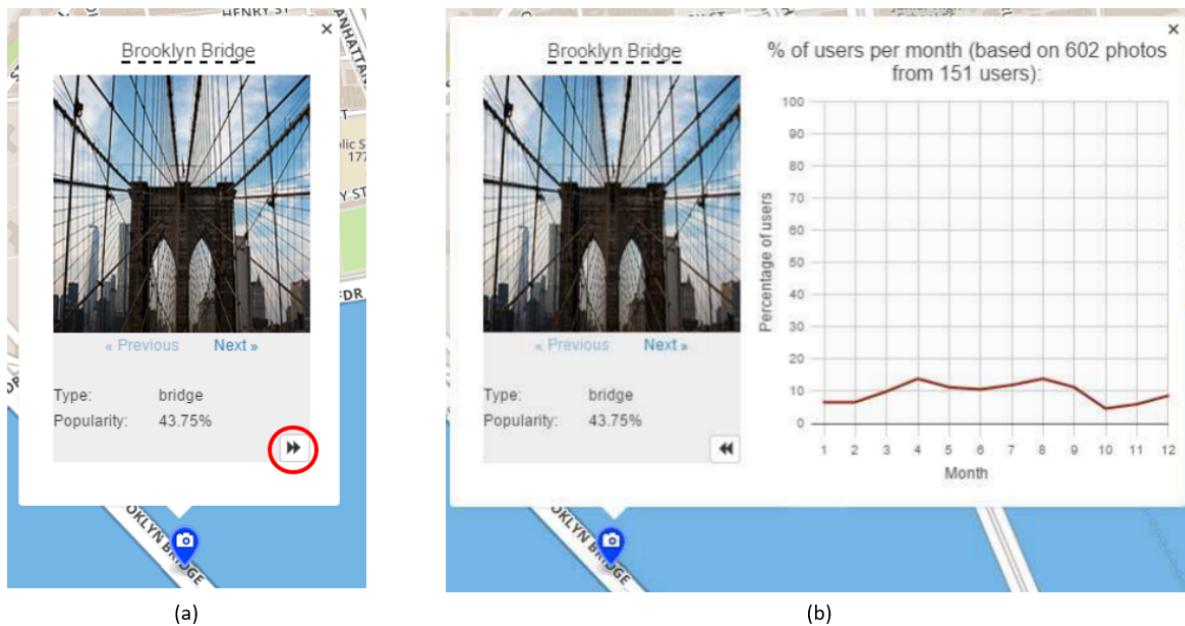


Figura 17 – Informações sobre a ponte do Brooklyn, Nova York, EUA: (a) informações básicas; (b) frequência de visitas realizadas ao longo de 2014 (funcionalidade acessível por meio do botão destacado pelo círculo vermelho em (a)).

Outra funcionalidade disponível apresenta estatísticas de visitas do PoI ao longo do ano anterior (Figura 17(b)). Assim, o usuário pode ser capaz de reconhecer PoIs sazonais e planejar a melhor época para visitar os locais desejados. A frequência de visitas de um PoI é obtida calculando-se a quantidade de usuários distintos que tiraram fotos naquele ponto turístico a cada mês, tendo como base as fotografias resultantes de uma consulta no Flickr utilizando o nome do local e o período correspondente ao ano anterior.

Outra funcionalidade diz respeito à quantidade e relevância dos PoIs que são exibidos. Opcionalmente, o usuário pode interagir com o mapa a fim de visualizar apenas os PoIs com popularidade igual ou superior a um valor de sua escolha. A Figura 18 apresenta duas filtragens de exibição de PoIs conforme a popularidade, uma com valor mínimo de 47% e outra com 78%. É possível ver que a quantidade de PoIs exibidos diminui na medida em que a popularidade mínima solicitada aumenta.

O sistema também possui integração com *Bing Knowledge Widget*, que identifica entidades na interface e recupera informações sobre elas quando solicitado. Assim, quando o usuário seleciona o nome de um PoI, uma tela na lateral do sistema mostra diversas informações relacionadas a ele, as quais representam os resultados da busca realizada por esse serviço. A Figura 19 apresenta um exemplo desse recurso, a partir do qual o usuário pode aprender ainda mais a respeito do PoI.

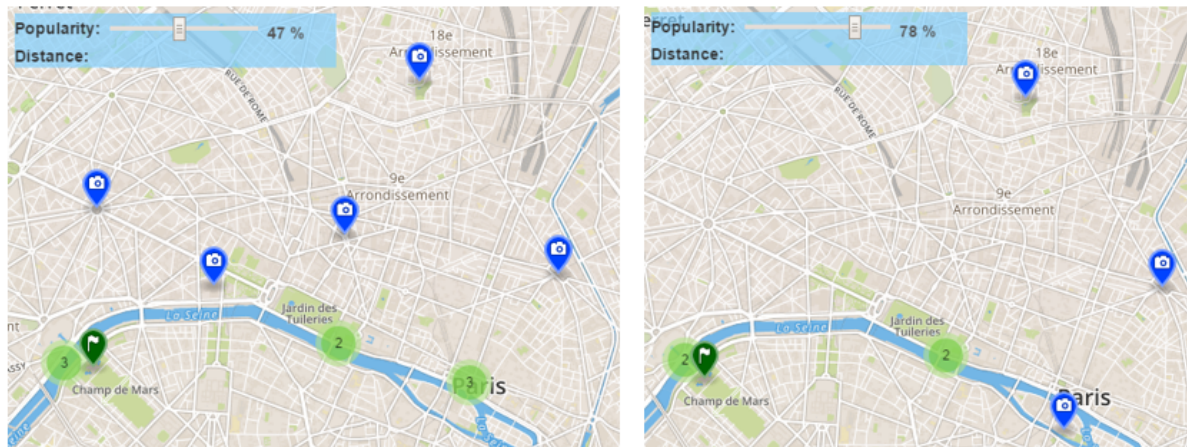


Figura 18 – Exibição de PoIs no mapa conforme a popularidade mínima escolhida pelo usuário.

Quando o usuário informa o local desejado para ser o centro da busca, o sistema obtém e exibe a estrutura hierárquica da localização em termos de cidade, estado e país. Assim, por meio do *Bing Knowledge Widget*, o usuário também pode obter informações e aprender sobre a cidade turística e país de destino. A Figura 20 apresenta alguns exemplos do reconhecimento das cidades.

É possível que o *Bing Knowledge Widget* não reconheça como entidades todas as cidades ou PoIs identificados. A Figura 21 mostra um exemplo no qual o PoI *Place de la République* foi reconhecido (e, por isso, é apresentado com sublinhado no nome) e o *The Louvre* não. Nesse caso, o usuário pode selecionar o nome do PoI não reconhecido e solicitar uma busca por meio do botão que é exibido. Em ambos os casos, a tela lateral surge com as informações encontradas.

Por fim, quando solicitado pelo usuário, o sistema calcula e exibe dois tipos de distância: a distância Haversine entre a localização central da busca e o PoI sendo analisado; e a distância Haversine entre quaisquer dois PoIs de interesse do usuário. Esse recurso visa auxiliar o turista em seu planejamento de forma que ele possa, por exemplo, utilizar a distância entre os PoIs para elaborar a sequência de visitas.

4.3 Execução em tempo real

Para tirar proveito da dinamicidade do repositório *online* de fotografias, é importante que todo o processo seja executado em tempo real, da obtenção das fotografias no repositório até a recuperação das informações dos PoIs identificados a partir dessas fotos.

A API do Flickr, apesar de pública e gratuita, tem como limitação o retorno de, no máximo, 250 imagens para cada requisição, o que faz com que seja necessário realizar

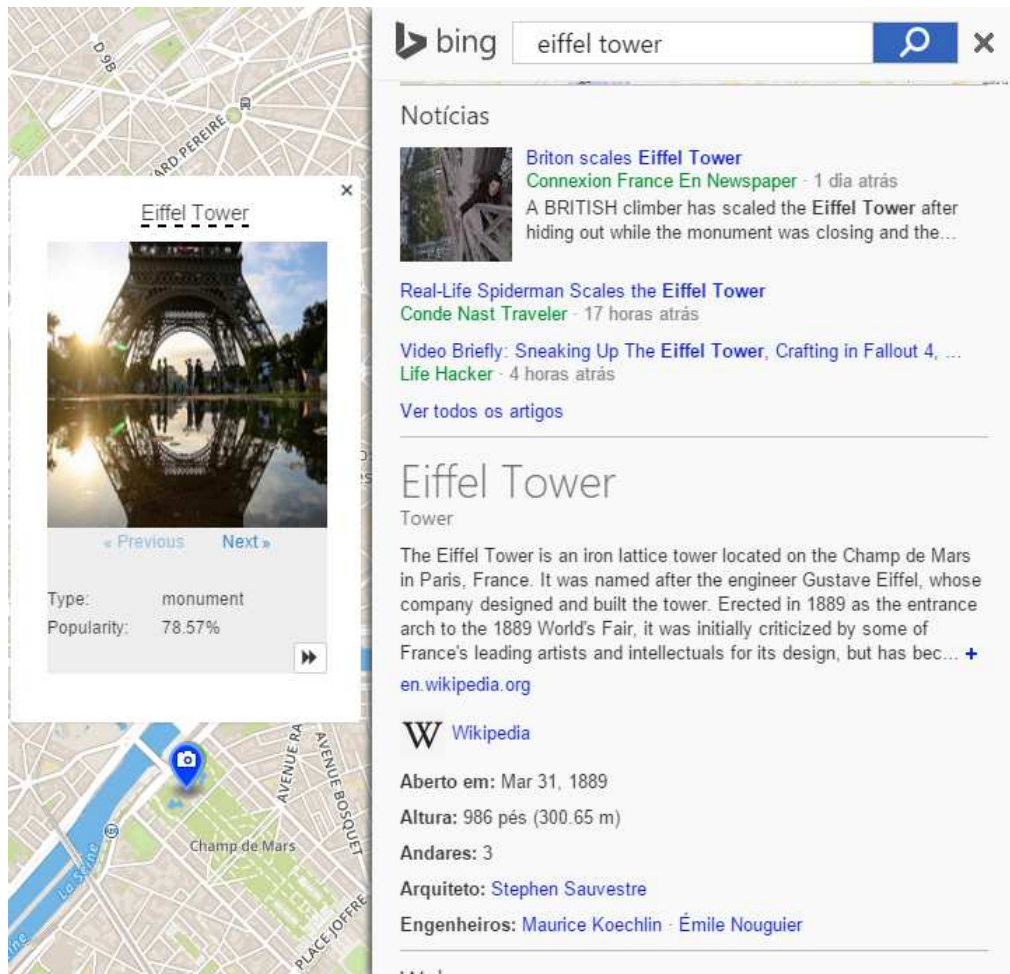


Figura 19 – Informações sobre a Torre Eiffel (Paris, França) retornadas e exibidas pelo *Bing Knowledge Widget*.



Figura 20 – Exemplos de cidades onde encontram-se as localizações pesquisadas.

várias solicitações. Para contornar a demanda de tempo exigida por solicitações sucessivas, o sistema *ToPI Trip* implementa uma solução *multi-thread*.

O usuário pode não ter interesse em explorar todos os PoIs identificados. Portanto, para evitar desperdício de processamento e garantir melhor tempo de resposta, o sistema foi implementado de acordo com os três fluxos apresentados na Figura 22. Assim, as consultas às fontes de informação (por palavras-chave e, eventualmente, por proximidade

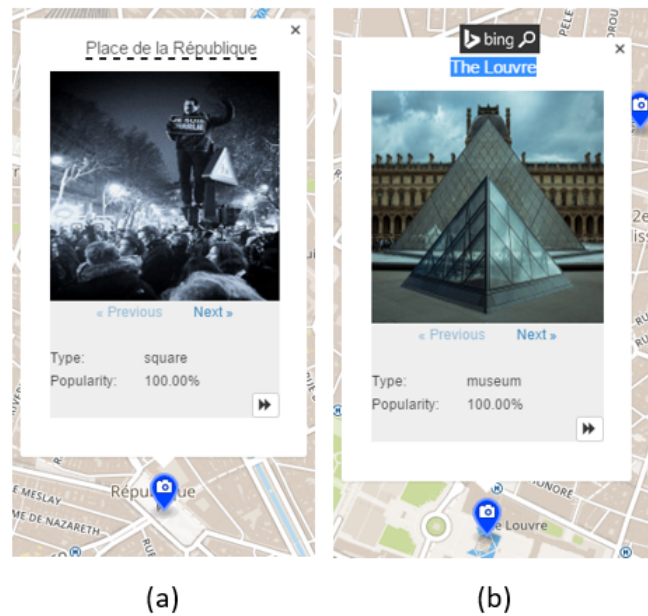


Figura 21 – Reconhecimento de PoIs pelo *Bing Knowledge Widget*: (a) PoI reconhecido - por isso há o sublinhado no nome; (b) PoI não reconhecido (sem sublinhado) - nesse caso, o usuário pode selecionar o nome e solicitar uma busca por meio do botão “bing”.

geográfica) são realizadas para um determinado PoI apenas quando o usuário o seleciona na interface (2º fluxo). Isso é importante para evitar que haja desperdício de tempo e processamento trazendo informações sobre todos os PoIs retornados sem que o usuário tenha interesse em ver tais informações. Da mesma forma, pode ser que o usuário não queira visualizar a frequência de visitas do local e, por isso, essa informação também é obtida sob demanda (3º fluxo).

4.4 Configuração do sistema

Alguns módulos do método ToPI exigem a escolha prática de um serviço ou abordagem para que sejam executados. É o caso da escolha do repositório *online* de fotografias, da técnica de pontuação de *tags* e do algoritmo de agrupamento. Esta seção descreve os critérios utilizados na configuração dessas três etapas. A Seção 4.4.1 apresenta alguns repositórios de fotografias disponíveis e justifica a escolha do Flickr. A Seção 4.4.2 apresenta a técnica de pontuação utilizada neste trabalho a fim de obter as palavras-chave que serão utilizadas para descrever os PoIs. Por fim, a Seção 4.4.3 apresenta os critérios e testes considerados na escolha do método de agrupamento que será aplicado na identificação dos PoIs.

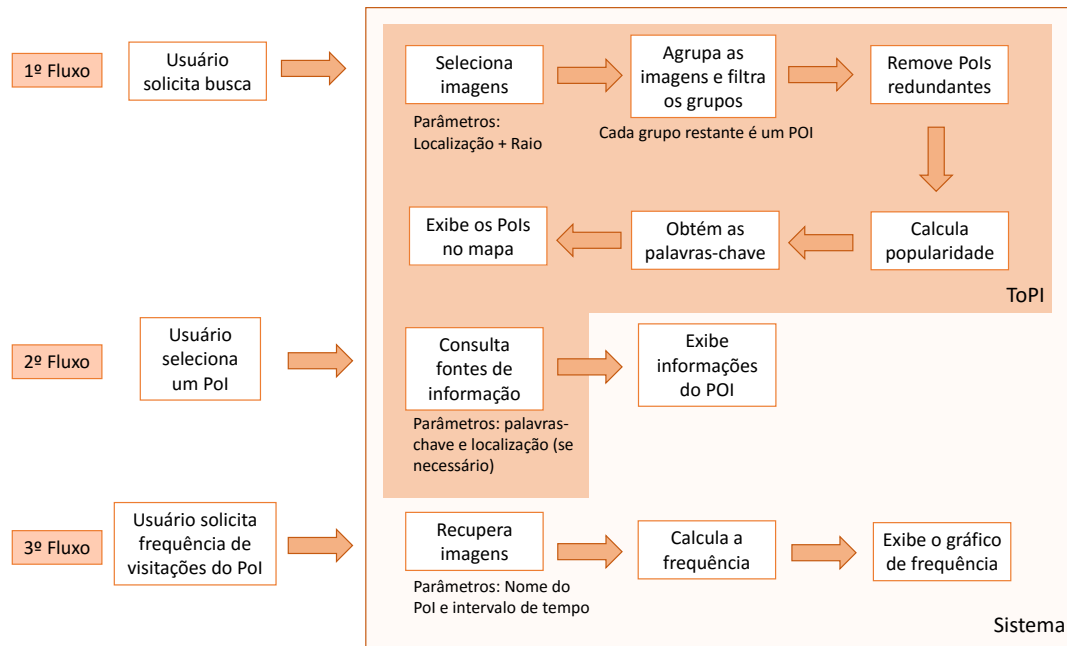


Figura 22 – Sequência de fluxos executados pelo sistema.

4.4.1 Escolha do repositório *online* de fotografias

Uma das redes sociais focadas no contexto de imagens é o Instagram², rede que contém mais de 30 bilhões de fotografias compartilhadas e possui uma média de 70 milhões de *uploads* diários (Instagram, 2015a). Apesar da alta quantidade de imagens, sua API impõe duas restrições que limitam sua utilização para o propósito de identificação de POIs (Instagram, 2015b): a consulta por fotografias em uma região geográfica deve respeitar um limite espacial (no máximo 5 km para o raio que define essa região) e um temporal (apenas as fotos tiradas em um período máximo de sete dias são recuperadas).

Outra plataforma social é o Panoramio³, serviço pertencente ao Google que possui 80 milhões de fotos publicadas, mas que está na iminência de ser encerrado (CNET, 2015). Assim como o Instagram, a utilização desse repositório também mostrou-se inviável para o propósito deste trabalho.

Já a rede social Flickr possui cerca de 5,26 bilhões de fotografias públicas, com cerca de 728 milhões postadas apenas em 2015 (o equivalente a cerca de 2 milhões de *uploads* por dia) (MICHEL, 2016). Embora possua menos fotos do que o Instagram, sua API é mais abrangente: possui limite de 32 km para o raio da área geográfica a ser pesquisada e não limita o intervalo de tempo a ser consultado (Flickr, 2015). A API também limita em 4000 a quantidade de imagens que pode ser retornada pela busca. Apesar disso, essa é uma quantidade aceitável para o funcionamento do método ToPI e não impede a escolha

² <https://www.instagram.com/>

³ <http://www.panoramio.com/>

do Flickr como repositório de imagens. A plataforma Flickr é comumente utilizada em diversos trabalhos da literatura, como apresentado na Seção 2.2, e também ilustra os experimentos realizados neste trabalho.

4.4.2 Método de pontuação das tags

A qualidade das palavras-chave que são utilizadas para descrever um PoI e obter mais informações sobre ele depende do critério utilizado na pontuação das *tags*. Se forem associadas pontuações elevadas às *tags* pouco relevantes, as palavras-chave identificadas podem não representar corretamente o local e seu uso poderia, por exemplo, resultar em informações erradas.

O sistema ToPI *Trip* adotou a estratégia de pontuação de *tags* proposta no trabalho (HE et al., 2012). Essa abordagem é apresentada a seguir.

Considerando t uma *tag*, C o conjunto de grupos resultantes do processo de agrupamento e $c \in C$ um desses grupos, F_c denota o conjunto de fotografias em c e n_t representa o número de grupos em C que contém t . Além disso, $U_{t,c}$ representa o conjunto de usuários distintos em c que utilizaram a *tag* t e $F_{t,c}$ refere-se ao conjunto de fotos em c que contém t .

A pontuação de uma *tag* t em c é definida como:

$$Pontuacao(t, c) = \frac{(tf(t, c) \cdot idf(t) \cdot uf(t, c))}{mtf(t, c)}$$

Onde

$$tf(t, c) = |F_{t,c}|/|F_c| \qquad idf(t) = \log(|C|/n_t)$$

$$uf(t, c) = |U_{t,c}| \qquad mtf(t, c) = \sum_{c \in C} tf(t, c)/|C|$$

O método de pontuação acima é uma versão derivada da clássica medida estatística TF-IDF (SALTON; BUCKLEY, 1988), muito utilizada na área de Recuperação da Informação com o objetivo de indicar a importância de um termo em relação a um conjunto de documentos. Na versão clássica, o componente tf refere-se à frequência de um termo t em um documento específico e o componente idf indica quão importante (ou específico) o termo é. No contexto deste trabalho, cada termo é uma *tag* e cada documento refere-se a um grupo. Trata-se, portanto, de um cálculo que leva em consideração, além da adequação do tradicional TF-IDF ao contexto de *tags*, a proporção de usuários do grupo que utilizaram uma dada *tag* (componente uf) e sua distribuição geográfica (componente mtf).

4.4.3 Método de agrupamento

Para avaliar as técnicas de agrupamento, duas bases de imagens foram criadas. A primeira contém fotos de 10 PoIs populares em Paris, França e foi chamada de DB_P. A segunda possui imagens de 10 locais em Nova York, nos Estados Unidos e recebeu o nome de DB_{NY}. Para a criação dessas bases, o Flickr foi consultado utilizando como chave de busca o nome de cada PoI e a lista de fotos retornadas foi ordenada de acordo com o algoritmo de relevância provido pela API. Para cada local, no máximo 150 fotografias foram recuperadas com a condição de que a distância de cada imagem ao PoI em questão fosse igual ou menor a 50 metros. As coordenadas GPS dos PoIs foram obtidas em *MyGeoPosition.com*. A Tabela 3 lista cada PoI considerado e seu correspondente número de fotografias.

Tabela 3 – PoIs utilizados como *ground truth*.

Paris		Nova York	
PoI	Número de fotos	PoI	Número de fotos
Arc de Triomphe	150	Grand Central Terminal	127
Orsay Museum	21	Empire State Building	117
Eiffel Tower	108	Washington Square Park	58
Musee du quai Branly	23	New York Public Library	55
Grand Palais	150	One World Trade Center	56
Louvre Museum	88	Rockefeller Center	40
Notre Dame Cathedral	47	Statue of Liberty	150
Pantheon	27	The Metropolitan Museum of Art	120
Petit Palais	54	The Museum of Modern Art	44
Place des Vosges	25	Times Square	73

Foram testados todos os métodos de agrupamento baseados em densidade apresentados na Seção 2.1.2 (DBSCAN, OPTICS, *Mean Shift*, HDBSCAN e P-DBSCAN) usando as duas bases de dados criadas. A medida *F-Measure* (Equação 1), por considerar precisão e revocação na análise da acurácia, foi usada para avaliar a qualidade de cada agrupamento.

$$F\text{-Measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

sendo que

$$Precision = \frac{tp}{tp + fp} \quad \quad \quad Recall = \frac{tp}{tp + fn}$$

onde *tp* (*true positive*) representa a quantidade de PoIs que foram identificados e pertencem ao *ground truth*; *fp* (*false positive*) representa o número de PoIs identificados

Tabela 4 – Comparação dos métodos de agrupamento baseados em densidade utilizando *F-Measure*. Quanto mais próximo do valor 1 (um), melhor.

	DBSCAN	OPTICS	P-DBSCAN	HDBSCAN	Mean Shift
DB _P	1	1	0,75	0,823	0,947
DB _{NY}	1	1	0,889	0,889	1

incorretamente (que não pertencem ao *ground truth*); e *fn* (*false negative*) simboliza a quantidade de PoIs pertencentes ao *ground truth* e que não foram identificados.

Os parâmetros utilizados para agrupar as fotos são $MinPts = 5$ e $Eps = 20$ metros (nem todos os algoritmos requerem ambos os parâmetros) e a medida de distância adotada é a Haversine. Para o método *Mean Shift*, uma foto é considerada como vizinha se sua distância Haversine à foto sendo analisada não é maior que Eps . O método P-DBSCAN foi aplicado com *adaptive density threshold* igual a 10%, tal como utilizado pelos autores do método em seus experimentos (KISILEVICH; MANSMANN; KEIM, 2010).

O sistema foi executado por duas vezes para cada método de agrupamento, uma vez utilizando as imagens pertencentes à base DB_P e a outra utilizando DB_{NY}. Grupos com poucos usuários foram descartados ($T_{users} = 3$) e remoções de PoIs redundantes foram realizadas quando necessário ($T_{MergeD} = 50$ metros e $T_{MergeK} = 200$ metros). A quantidade de palavras-chave utilizada nas remoções de redundâncias e nas consultas por informações foi definida como 10 ($N_{pc} = 10$), mas percebeu-se, ao fim destes experimentos, que todos os PoIs de DB_P e DB_{NY} foram rotulados com, no máximo, 4 palavras-chave. Por essa razão, o valor escolhido para ser executado por padrão no sistema e nos experimentos do capítulo 5 foi $N_{pc} = 4$.

A avaliação dos algoritmos de agrupamento foi realizada em um computador com 4GB de memória RAM e processador Core 2 Duo 2.20 GHz. A fonte de informações utilizada na identificação dos PoIs pelo sistema foi o GeoNames⁴, base de dados geográfica, gratuita e colaborativa, que fornece, dentre outras informações, o nome do local, suas coordenadas geográficas e o endereço da respectiva página na Wikipedia. Uma grande vantagem do GeoNames é que ele conhece nomes alternativos (em outros idiomas) dos locais e, por isso, é capaz de obter o resultado da busca independente do idioma da palavra-chave.

Conforme mostra a Tabela 4, os algoritmos com melhores resultados foram DBSCAN e OPTICS. A Tabela 5 apresenta o tempo médio de execução gasto por cada método⁵. Como pode ser notado, o algoritmo OPTICS aliou o menor tempo de execução ao reconhecimento correto de todos os PoIs testados. Por essa razão, ele foi escolhido para ser utilizado no sistema.

⁴ <http://www.geonames.org/>

⁵ Tempo médio calculado a partir de três execuções sucessivas. O tempo de execução considera o tempo demandado tanto pelo agrupamento quanto pela filtragem de acordo com o número de usuários distintos nos grupos. O algoritmo P-DBSCAN considera a quantidade mínima de usuários no processo de agrupamento; nos outros métodos, essa filtragem é realizada na forma de pós-processamento.

Tabela 5 – Comparação dos métodos de agrupamento em relação ao tempo médio de execução (em segundos).

	DBSCAN	OPTICS	P-DBSCAN	HDBSCAN	Mean Shift
DB _P	0,524	0,515	0,625	0,786	1,250
DB _{NY}	0,613	0,579	0,804	0,749	1,386

4.5 Considerações Finais

Este capítulo apresentou o sistema ToPI *Trip*, desenvolvido para avaliar o método ToPI apresentado no capítulo anterior. Vários recursos de interação foram incorporados ao sistema visando facilitar e acelerar a obtenção de informações e tomadas de decisão. Por meio da integração com o serviço *Bing Knowledge Widget*, o usuário visualiza detalhes e aprende mais sobre o PoI sob análise sem que tenha que deixar o sistema e procurar por outras fontes de informação.

Para tirar proveito da dinamicidade presente nos repositórios de fotografias e proporcionar uma melhor experiência ao usuário quanto ao tempo de resposta, o sistema foi criado utilizando técnicas e fluxos que permitam sua execução em tempo real.

Por fim, o Flickr foi escolhido como repositório de fotografias e as abordagens utilizadas em duas etapas importantes do processo foram apresentadas: a etapa de pontuação das *tags*, primeira das duas, que é realizada utilizando a proposta desenvolvida em (HE et al., 2012), e o agrupamento das imagens, que é realizado pelo método OPTICS - escolhido após testes com outras quatro técnicas de agrupamento baseadas em densidade.

Experimentos e Análise dos Resultados

Este capítulo descreve os experimentos que demonstram a qualidade alcançada com a utilização do método ToPI em diversos cenários. A Seção 5.1 apresenta experimentos que analisam a relevância dos PoIs identificados. A Seção 5.2 traz alguns exemplos do funcionamento do método e do sistema em contexto dinâmico. A Seção 5.3 descreve experimentos que demonstram a importância da identificação de palavras-chave ao invés do uso direto das *tags* com maiores pontuações. A Seção 5.4 mostra o tempo médio de execução do sistema em cada etapa do método ToPI apresentada na Figura 9. Por fim, a Seção 5.5 encerra o capítulo com as considerações finais.

5.1 Relevância dos PoIs identificados

5.1.1 Método para a Avaliação

A avaliação da qualidade do método em relação à relevância dos PoIs identificados considera a taxa de concordância entre os PoIs identificados pela proposta deste trabalho e o TripAdvisor¹, um dos mais populares *websites* de turismo no mundo. Além disso, também é considerado quão consistentes os PoIs encontrados são de acordo com o Google Maps², em termos de localização geográfica.

A base de dados de comparação utilizada em cada experimento consiste de uma lista de PoIs do TripAdvisor formada por locais pertencentes a quatro subcategorias da categoria “O que fazer em” na cidade do experimento, sendo elas: “Pontos turísticos e de interesse”, “Museus”, “Natureza e parques” e “Teatros e shows”. A lista foi filtrada de modo que contenha apenas locais que possuam coordenadas GPS conhecidas pelo *website* e que estejam dentro da área geográfica definida pela localização central e pelo raio de abrangência da busca.

¹ <http://www.tripadvisor.com>

² <https://www.google.com/maps>

Tabela 6 – Valores dos parâmetros utilizados nos experimentos.

Etapa	Parâmetro	Valor
Seleção das imagens	Quantidade requisitada (n_{max})	3000
	Raio de abrangência	10 km
Agrupamento e filtragem	Método de agrupamento	OPTICS
	Medida de distância	Haversine
	Eps	50 metros
	$MinPts$	6
	T_{users}	3
Remoção de PoIs redundantes	T_{MergeD}	50 metros
	T_{MergeK}	200 metros
	N_{pc}	4
Obtenção das informações	Fonte de informações	GeoNames
	$T_{maxdist}$	500 metros

Tabela 7 – Cidades pesquisadas nos experimentos.

Cidade	Centro da busca			Qtd imagens
	Local	Latitude	Longitude	
Paris, França	Torre Eiffel	48.858312	2.29449	2.295
Rio de Janeiro, Brasil	Cristo Redentor	-22.951981	-43.211009	2.297
Nova York, EUA	Empire State Building	40.748440	-73.985655	2.959

A Tabela 6 detalha os valores empíricos utilizados nos parâmetros dos experimentos realizados. A Tabela 7 apresenta as cidades e localizações utilizadas como centro das consultas, bem como a quantidade de imagens efetivamente utilizada em cada uma. Cabe destacar que todas as imagens recuperadas estão sob licenças que permitem redistribuição³ e foram tiradas no período de Setembro/2014 a Setembro/2015.

Considera-se que há concordância na identificação de um PoI se ao menos um dos critérios abaixo é atendido (caso mais de um critério seja satisfeito, apenas o primeiro deles é considerado nos resultados):

- ❑ C1: O nome do PoI identificado é similar ao nome de um dos três locais do TripAdvisor mais próximos a ele segundo a distância de Haversine;
- ❑ C2: Há um PoI do TripAdvisor próximo ao local identificado pelo método, isto é, com distância de Haversine igual ou inferior a 50 metros;
- ❑ C3: Não há concordância com o TripAdvisor, porém o Google Maps indica que o PoI identificado foi plotado corretamente no mapa.

³ <https://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html>

O primeiro critério considera três locais do TripAdvisor para reduzir possíveis distrações, como, por exemplo, uma estátua em frente ao PoI. A segunda condição, por sua vez, é capaz de obter concordância mesmo se não houver nomes similares (devido à diferença de idiomas, por exemplo). Além disso, possibilita que haja concordância mesmo quando o método não consegue rotular um determinado local. Neste caso, como há um PoI muito próximo dele no TripAdvisor, considera-se que há concordância, já que tal local, mesmo não rotulado, pode ser de interesse do usuário. Por fim, o terceiro critério é importante pois atende lugares extensos, como em situações nas quais a latitude e longitude do TripAdvisor estão situadas em uma posição específica do local e o PoI identificado pelo método está situado em uma posição diferente, também dentro do local. Esse último critério também atende a situação “lugar vs evento”: como não é feita distinção entre um lugar e um evento que ocorre nele, o método é considerado correto se retornar qualquer um dos dois. Por exemplo, considera-se válida a identificação de um PoI p tanto como *Javits Center*, em Nova York, quanto como *Comic Con* (evento sediado nesse local). Nessas situações, o usuário pode utilizar as funcionalidades apresentadas na Seção 4.2 para efetuar uma análise mais detalhada sobre o PoI e evitar enganos. Outro exemplo dessa situação pode ser visto na Figura 24, onde um PoI identificado em Paris foi rotulado como um evento (*Paris Games Week*) ao invés de seu local (*Paris expo Porte de Versailles*).

5.1.2 Experimentos

A Tabela 8 apresenta os resultados obtidos, detalhando a quantidade de PoIs identificados por cada palavra-chave (PC_x), a quantidade reconhecida por proximidade geográfica e a quantidade de concordâncias, considerando cada critério (C_x).

Tabela 8 – Resultados dos experimentos. PC_x indica qual das quatro palavras-chave foi usada na identificação do PoI. C_x refere-se ao critério utilizado para que houvesse concordância.

Cidade	PoIs identificados						PoIs em concordância			
	PC_1	PC_2	PC_3	PC_4	Proximidade	Total	C_1	C_2	C_3	Total
Paris	11	4	2	0	11	28	17	5	5	27
Rio de Janeiro	1	2	1	0	5	9	8	0	1	9
Nova York	5	4	0	1	13	23	14	3	4	21

5.1.2.1 Centro da busca: Torre Eiffel - Paris

O primeiro experimento consistiu de uma busca por PoIs próximos à Torre Eiffel. A maioria dos locais identificados foi reconhecida por meio de palavras-chave, especialmente pela primeira delas (11 de 17 PoIs), característica que valida a estratégia de pontuação das *tags* e obtenção das palavras-chave.

A Figura 23 apresenta quatro PoIs identificados nesse experimento. A presença de um parênteses duplo no nome do PoI *Grand Palais* indica que esse local foi identificado por meio de uma busca por proximidade geográfica e seu valor corresponde à distância (em metros) entre a localização do PoI calculada pelo método e a localização dele conhecida pelo GeoNames.

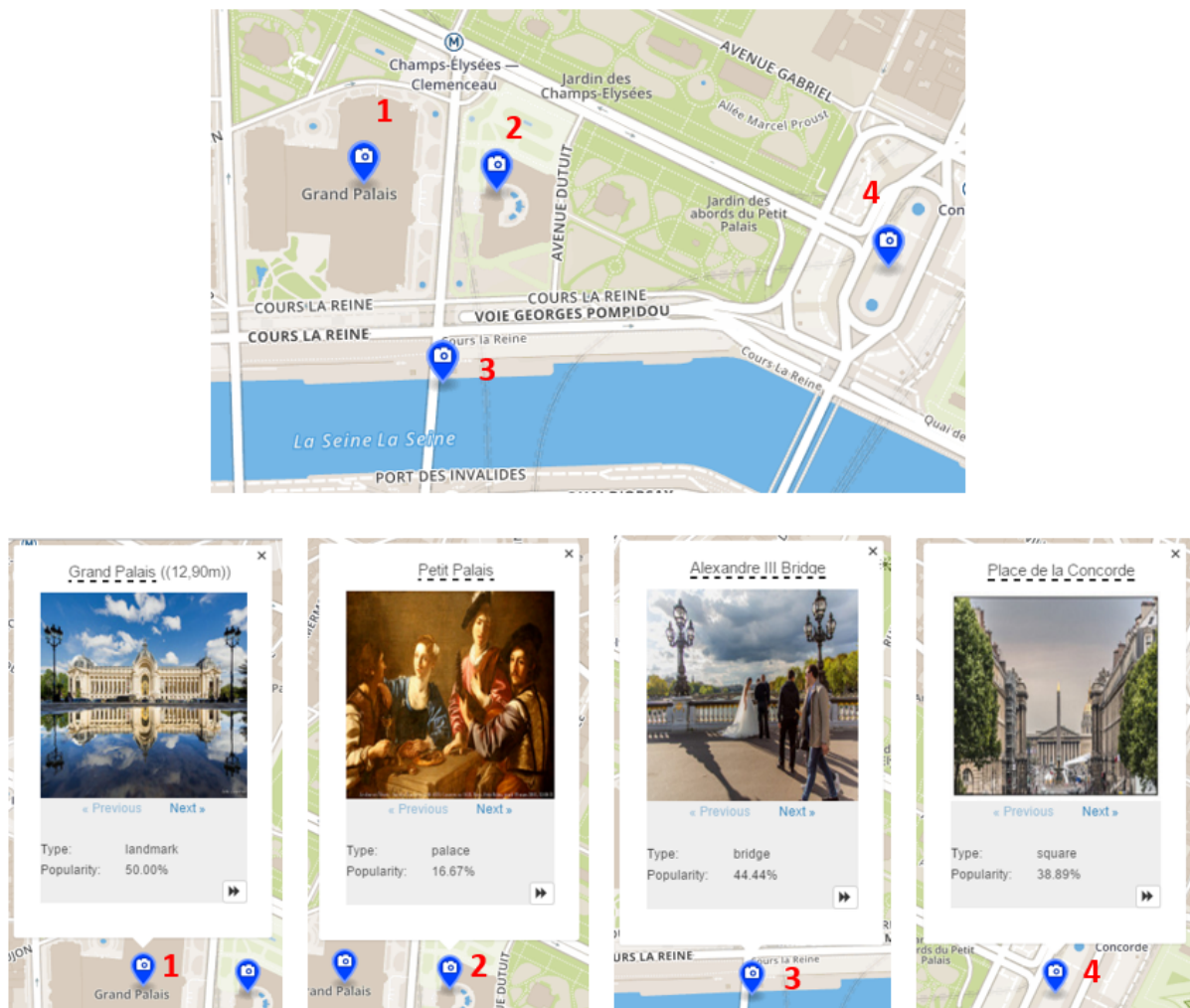


Figura 23 – Alguns PoIs retornados na busca em Paris.

Nesse experimento, uma das concordâncias obtidas é um caso da situação “lugar vs evento” no qual o PoI foi rotulado como *Paris Games Week*, um evento anual realizado no *Paris expo Porte de Versailles* (Figura 24). O único PoI cuja identificação foi considerada errada deveu-se ao fato de estar situado em um local que não é popular por si mesmo, mas, por ser um lugar entre os locais famosos *Square Rene Viviani* e *Notre-Dame Cathedral*, é utilizado por muitas pessoas para tirar fotos desses dois lugares (Figura 25). Assim, como a latitude e longitude pertencem ao local onde as fotos foram tiradas, e não ao “local-alvo” delas, um grupo foi formado lá. Apesar disso, essa situação demonstra a

utilidade do método na identificação não apenas de PoIs, mas, eventualmente, de lugares que proporcionam boas visões deles e que também podem ser visitados pelo turista.

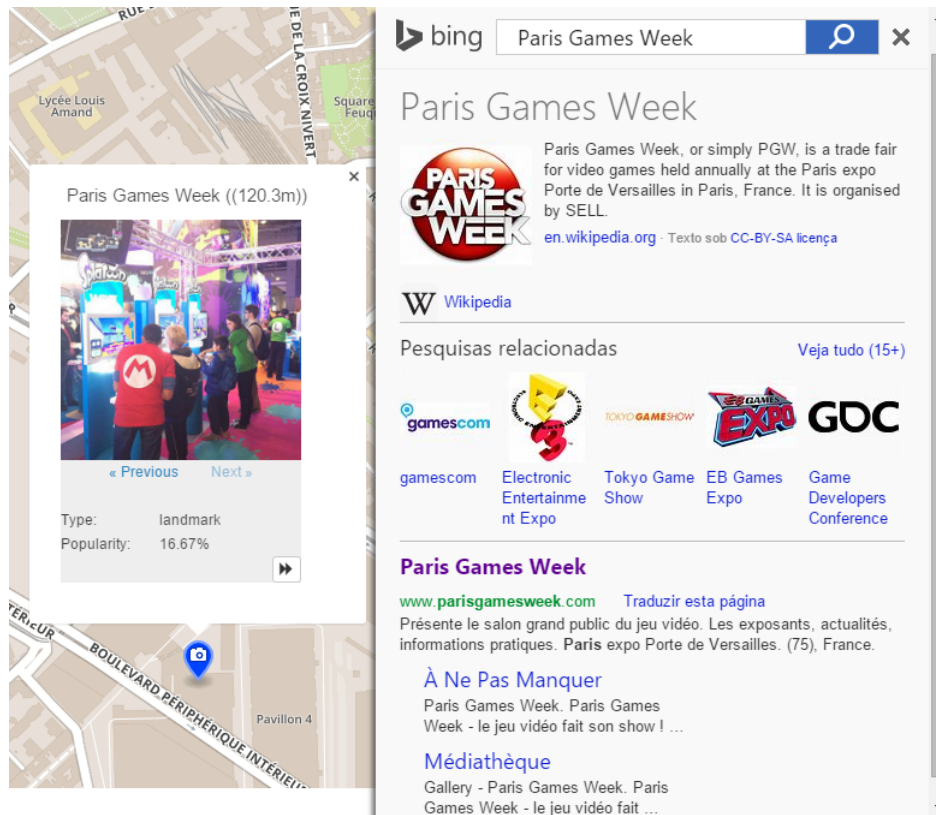


Figura 24 – Situação “lugar vs evento” no qual o PoI foi identificado como *Paris Games Week*, um evento anual sediado no *Paris expo Porte de Versailles*.

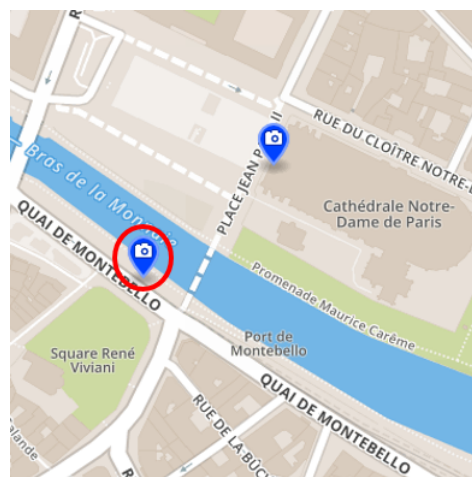
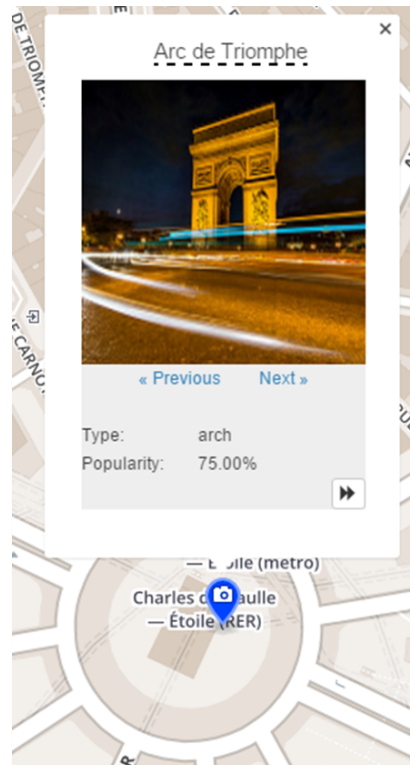


Figura 25 – PoI identificado incorretamente. O PoI destacado pelo círculo vermelho não é popular por si mesmo, mas sim um local onde as pessoas tiram fotografias dos locais *Square Rene Viviani* e *Notre-Dame Cathedral* (também identificado como um PoI e representado pelo outro marcador na imagem).

Outra situação percebida nesse experimento ressalta a importância da tratativa de PoIs redundantes. O famoso ponto turístico *Arc de Triomphe* foi representado, sem o devido tratamento, por dois grupos (Figura 26(a)), o que poderia confundir o usuário e passar a impressão de que são dois lugares distintos ou que houve erro na identificação. A remoção de redundâncias contorna este problema, tratando cada local de interesse com uma representação única (Figura 26(b)).



(a)



(b)

Figura 26 – Identificação do PoI *Arc de Triomphe* sem tratativa de PoIs redundantes (a); com a eliminação de redundâncias (b).

5.1.2.2 Centro da busca: Cristo Redentor - Rio de Janeiro

O segundo experimento utilizou as coordenadas geográficas do Cristo Redentor em prol de identificar locais turísticos próximos a ele. Conforme mostra a Tabela 8, todos os PoIs identificados nesse experimento estiveram de acordo com os critérios de concordância. A Figura 27 apresenta a localização e as informações básicas de três dos nove PoIs identificados. Os PoIs 2 e 3 (*Rio de Janeiro Cathedral* e *Escadaria Selarón*, respectivamente) não possuem categoria turística conhecida pelo GeoNames e, por isso, essa informação não é exibida. Assim como ocorreu no primeiro experimento, se não houvesse a remoção de PoIs redundantes, o local *Sugarloaf Mountain* também seria representado por dois grupos, como mostra a Figura 28.

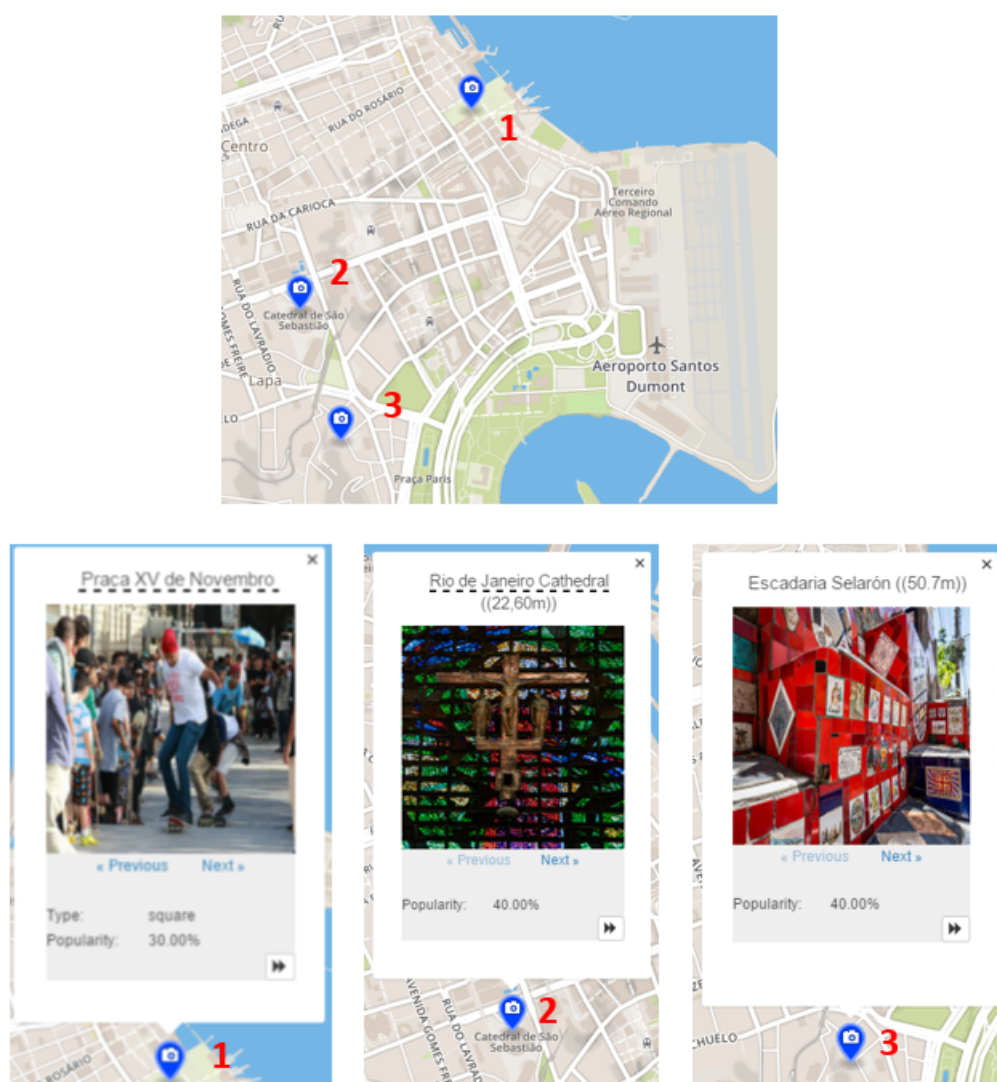


Figura 27 – Alguns PoIs retornados na busca no Rio de Janeiro.

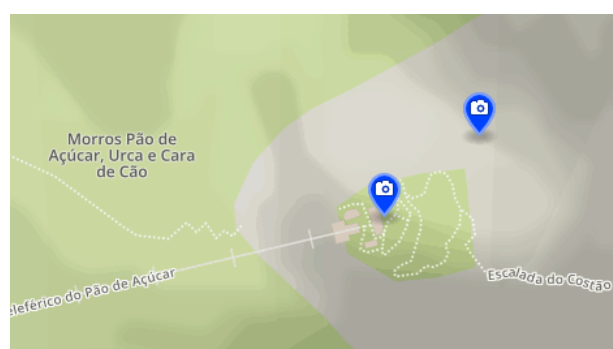


Figura 28 – Dois grupos representando o mesmo PoI (*Sugarloaf Mountain* ou Morro Pão de Açúcar, nome que aparece na imagem).

Quanto mais usuários do grupo utilizam uma determinada *tag*, maior a probabilidade de que ela seja mais relevante para o PoI e, assim, maior a chance de se tornar uma palavra-chave adequada para a consulta na fonte de informações. O PoI *Escadaria Selarón* é um exemplo que demonstra isso: apesar de possuir 10 fotografias pertencentes a quatro usuários distintos, apenas um deles inseriu tags para representar sua(s) foto(s); entretanto, além de poucas, as *tags* informadas são irrelevantes para a identificação do PoI (“*iphoneography*”, “*squareformat*”, “*instagramapp*”, “*square*” e “*rio*”). Como a busca por palavras-chave não obteve êxito, esse PoI só pôde ser identificado graças à busca por proximidade geográfica.

5.1.2.3 Centro da busca: *Empire State Building* - Nova York

O terceiro experimento identificou PoIs próximos ao *Empire State Building*, em Nova York. A Figura 29 apresenta quatro PoIs identificados.

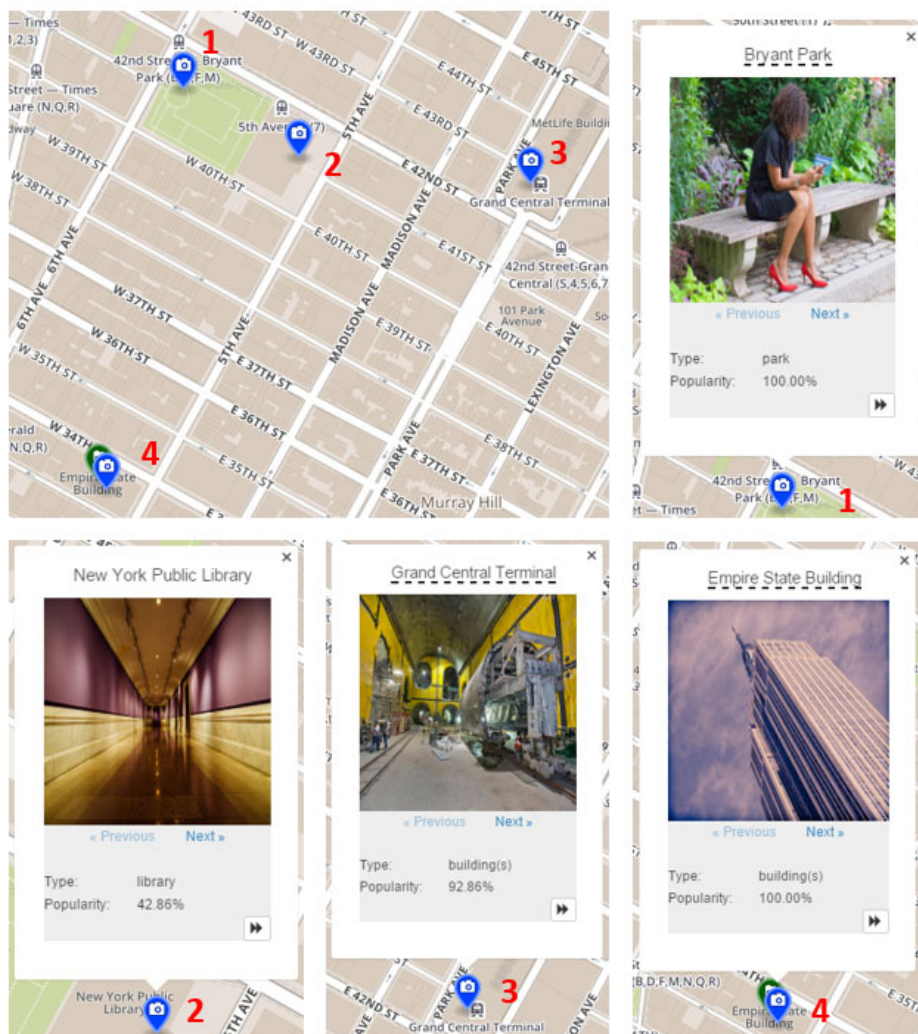


Figura 29 – Alguns PoIs retornados na busca em Nova York.

Não houve concordância em dois locais retornados nesse experimento, um deles representando uma situação especial. Nesse caso, o método identificou um lugar chamado *The Bushwick Collective*, mas o rotulou incorretamente devido ao fato de não haver tal local na base de dados do GeoNames (Figura 30). Como a abordagem é dependente da fonte de informações utilizada para rotular os PoIs, os resultados dela afetam fortemente as informações recuperadas pelo método.

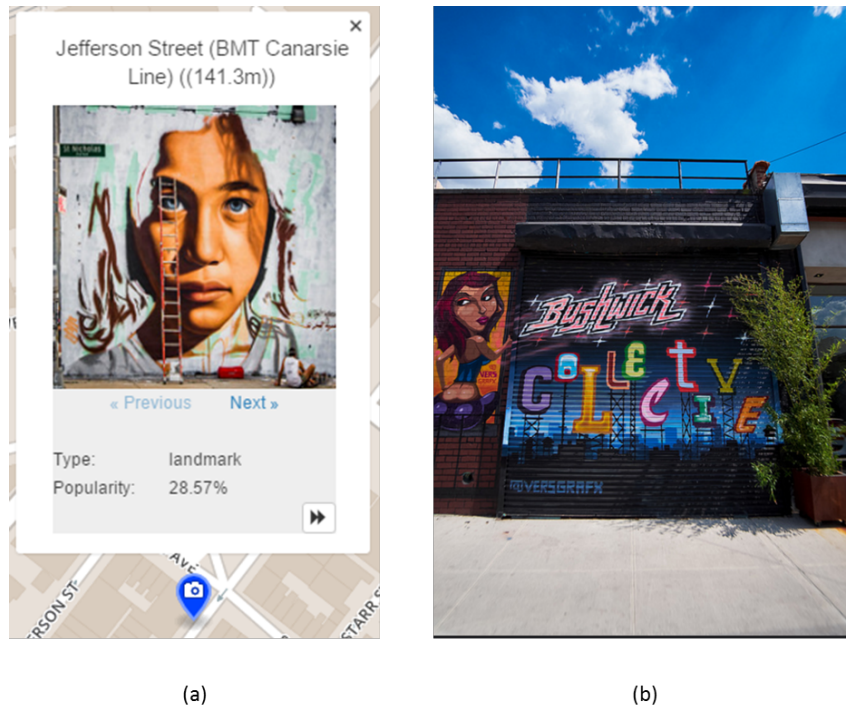


Figura 30 – Erro na obtenção de informações de um PoI: (a) PoI rotulado incorretamente (o correto seria *The Bushwick Collective*); (b) Fotografia pertencente ao PoI com o nome do local, reafirmando o erro na rotulagem.

O outro caso de não-concordância percebido nesse experimento não se refere a um caso especial. Nele, o método identificou um local como “*World Trade Center (PATH station)*”. No entanto, a análise dos três critérios para concordância não obteve êxito: nenhum dos três locais mais próximos que estão na base de comparação do TripAdvisor possui nome semelhante (critério 1); o local da base de comparação mais próximo a ele é o “*St. Paul’s Chapel*”, que está a 108 metros de distância dele (critério 2); e a localização do PoI em questão no Google Maps encontra-se a dois quarteirões de distância do local indicado pelo método (critério 3).

Adicionalmente à situação “lugar vs evento”, utilizar repositórios de imagens e metadados *online* para identificar PoIs pode levar a situações que podem ser confundidas com erros. Um exemplo é quando requisições consecutivas para o mesmo local e com os mesmos parâmetros retornam fotos diferentes e, assim, identificam locais diferentes. Situações como essa são basicamente causadas pela natureza dinâmica dos repositórios, na

qual imagens são incluídas e/ou excluídas no intervalo de tempo entre as duas consultas. Outra situação, que é consequência dessa, refere-se à presença de PoIs dentro de PoIs: uma consulta poderia resultar em um PoI identificado como um prédio e outra busca, com outras fotografias, poderia reconhecer o local como um restaurante dentro desse prédio. Um exemplo dessa situação pode ser visto na Figura 31, na qual um PoI foi identificado como *Rockefeller Center building* (em Nova York) em uma busca e como *Rainbow Room*, um restaurante dentro desse local, em outra.



Figura 31 – Situação “PoIs dentro de PoIs”: Uma busca retornou o edifício *Rockefeller Center* (imagem à esquerda) e outra retornou o *Rainbow Room*, restaurante dentro dele (imagem à direita).

5.2 Identificação de PoIs em contexto dinâmico

Esta seção apresenta alguns casos nos quais foram percebidos eventos que influenciaram a quantidade de imagens de alguns locais e, assim, permitiram que fossem retornados lugares que não eram recuperados em consultas feitas antes e/ou depois da ocorrência deles. Com exceção do raio de abrangência, todos os parâmetros dos experimentos dessa seção seguem os valores descritos na Tabela 6.

5.2.1 Atentados terroristas em Paris - nov/2015

Em 13 de novembro de 2015, Paris foi vítima de uma série de atentados terroristas, de autoria do Estado Islâmico, que deixou mais de 130 mortos (UOL, 2015). A repercussão

disso no Flickr fez com que o método ToPI fosse capaz de identificar dois PoIs que não eram recuperados em consultas realizadas antes disso. A Figura 32(a) apresenta uma região geográfica com apenas um PoI identificado a partir de fotografias tiradas entre 01/07/2015 e 01/10/2015, portanto, antes dos atentados. A Figura 32(b) apresenta os PoIs identificados nessa mesma região a partir de fotografias tiradas entre 01/10/2015 e 01/01/2016. O PoI 2, não rotulado pelo método, refere-se, segundo o Google Maps, ao *Café Bonne Bière*. Ele e o *Bataclan* (PoI 3) são dois dos locais atacados nos atentados. O centro utilizado nas buscas foi a localização do *Parc des Buttes-Chaumont* (latitude: 48.879449; longitude: 2.379999) e o raio adotado foi de 4 km.

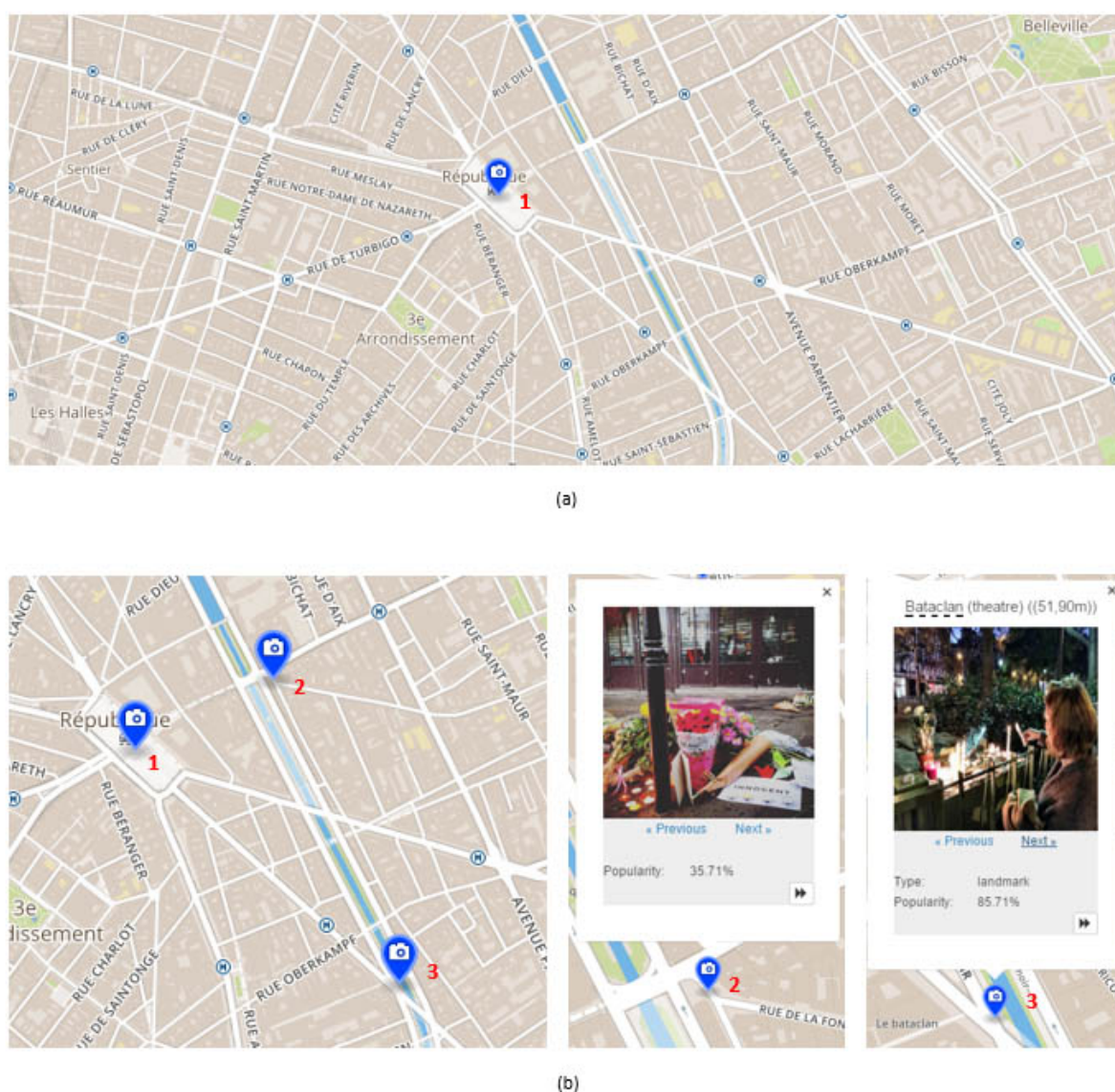
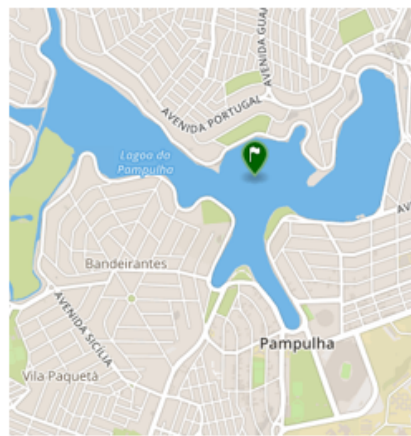


Figura 32 – Situação de dinamicidade do repositório de fotografias: (a) Consulta por PoIs realizada utilizando fotos tiradas antes dos atentados terroristas de 2015 em Paris; (b) Consulta utilizando fotos tiradas em um intervalo de tempo que cobre os atentados e consequente identificação de dois locais atacados.

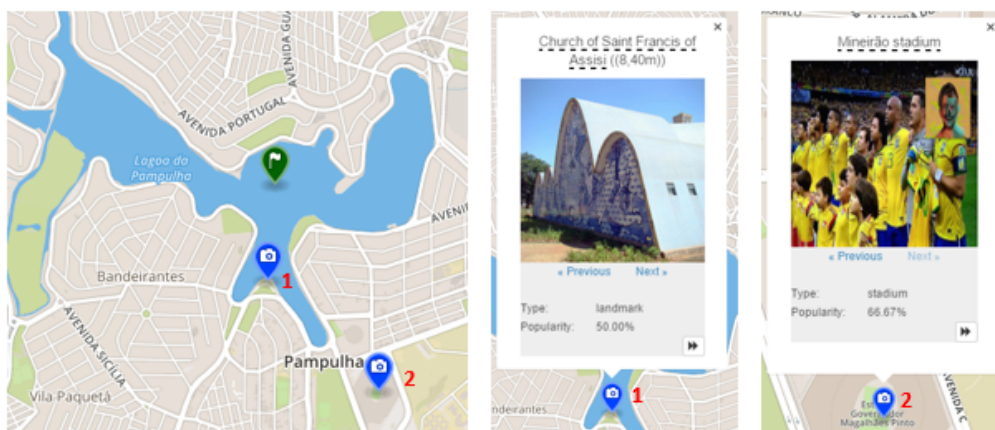
5.2.2 Copa do mundo de 2014

A copa do mundo de 2014, ocorrida no Brasil em junho e julho, também foi um evento que influenciou a identificação de alguns locais, especialmente aqueles que atraíram bastante movimento e que, por conta disso, destacaram-se no repositório utilizado.

O primeiro exemplo foi percebido em Belo Horizonte, Minas Gerais. Utilizando a Lagoa da Pampulha como centro de busca (latitude: -19.851735; longitude: -43.978528) e raio de 4 km, a Figura 33(a) mostra que nenhum PoI foi identificado na região apresentada quando pesquisado com fotos recentes (tiradas entre 09/02/2015 e 09/02/2016). Por outro lado, ao pesquisar utilizando fotos tiradas entre 01/01/2014 e 01/10/2014, é possível ver que houve a identificação dos PoIs Igreja São Francisco de Assis (*Church of Saint Francis of Assisi*) e o estádio Mineirão (Figura 33(b)). Apesar de não serem locais sazonais, o aumento de fotos dos locais devido à copa do mundo possibilitou a identificação.



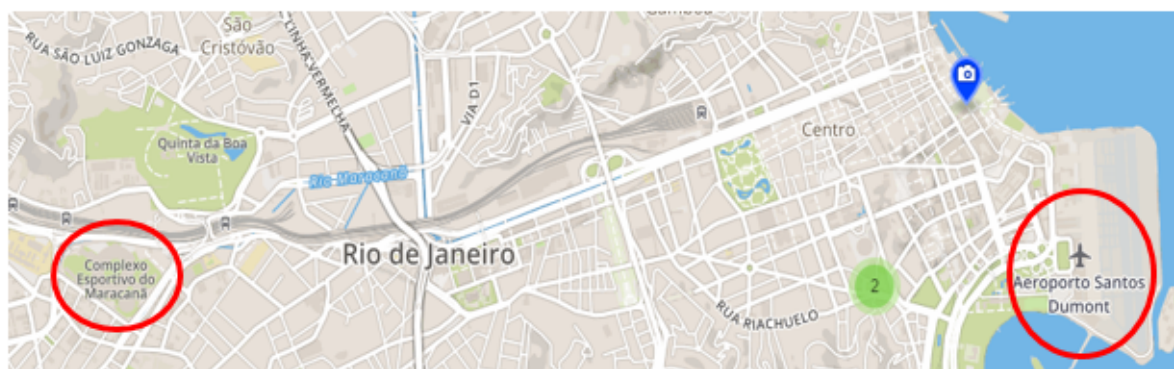
(a)



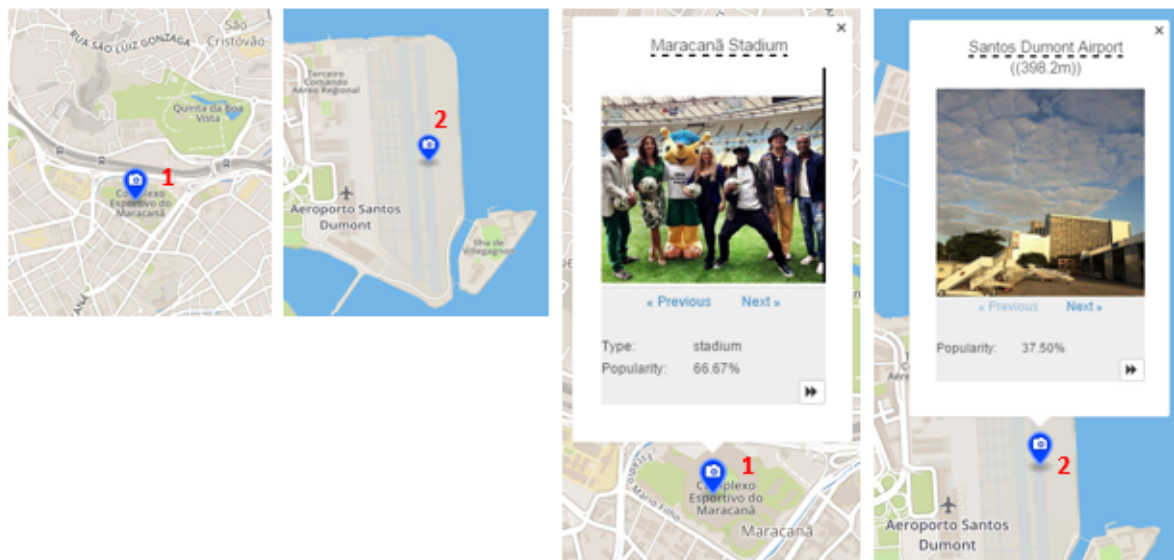
(b)

Figura 33 – Dinamicidade causada pela copa do Mundo 2014: (a) Consulta realizada após a realização do evento; (b) Consulta com fotos tiradas durante a copa, possibilitando a identificação dos dois locais apresentados.

Outro exemplo é a identificação do estádio do Maracanã e do aeroporto Santos Dumont, ambos no Rio de Janeiro. Como pode ser visto na Figura 34(a), a busca realizada com fotos tiradas entre 09/02/2015 e 09/02/2016 não identificou esses dois locais (destacados em vermelho). Os três PoIs identificados nessa busca são os mesmos apresentados na Figura 27. A Figura 34(b), por sua vez, mostra parte do resultado de uma busca utilizando fotografias tiradas entre 01/01/2014 e 01/10/2014. É possível ver que tanto o estádio quanto o aeroporto foram identificados. O centro utilizado nas buscas foi o Cristo Redentor (latitude: -22.951981; longitude: -43.211009), com raio de 10 km.



(a)



(b)

Figura 34 – Locais muito movimentados durante a copa do mundo de 2014: (a) Consulta realizada após a realização do evento - sem fotos suficientes para identificar o estádio do Maracanã e o aeroporto Santos Dumont; (b) Consulta com fotos tiradas durante a copa, possibilitando a identificação desses dois locais.

5.2.3 Museu do Amanhã - Rio de Janeiro

Um exemplo da identificação de PoIs recentes pode ser visto na Figura 35, que apresenta o *Museu do Amanhã* (*Museum of Tomorrow*). Esse museu foi inaugurado no Rio de Janeiro em 17 de dezembro de 2015 e já é identificado pelo método ToPI.

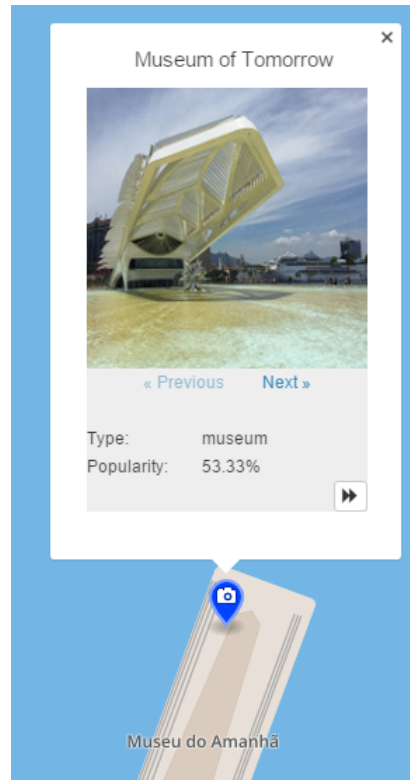


Figura 35 – Exemplo de identificação de PoIs recentes: *Museu do Amanhã* (*Museum of Tomorrow*), inaugurado em 17 de dezembro de 2015 no Rio de Janeiro.

5.2.4 Grande prêmio de Mônaco de Fórmula 1 - 2015

Outro evento que influenciou a quantidade de fotos de um determinado local no repositório foi o Grande prêmio (GP) de Mônaco, ocorrido entre 21 e 24 de maio de 2015. A Figura 36(a) mostra que nenhum PoI foi identificado na região apresentada, sendo exibido apenas o marcador referente ao centro da busca. Nessa consulta, foram utilizadas fotos tiradas entre 01/01/2015 e 01/05/2015, portanto, antes do evento. A Figura 36(b) apresenta o resultado de uma busca utilizando fotos do período de 01/05/2015 a 01/12/2015. É possível ver a que houve a identificação do Circuito de Mônaco, utilizado nas corridas. Como mostra o gráfico com estatísticas de visita desse local em 2015, maio é o mês mais visitado, o que coincide com o período do evento. O centro utilizado nas buscas está localizado no *Parc Princesse Antoinette* (latitude: 43.733460; longitude: 7.414762) e o raio considerado foi de 4 km.



(a)



(b)

Figura 36 – Influência de um evento na identificação de PoIs: (a) Consulta realizada após a realização do GP de Mônaco 2015 - nenhum PoI identificado na região da imagem; (b) Consulta com fotos tiradas no período do evento, possibilitando a identificação de um PoI.

5.3 Importância da utilização de palavras-chave

A fim de validar a estratégia de uso de palavras-chave em relação à utilização direta das *tags* com as maiores pontuações, esta seção apresenta um experimento que mostra a

Tabela 9 – Comparativo entre a quantidade de tentativas gastas na obtenção de informações utilizando (1) palavras-chave e (2) as tags com as maiores pontuações.

PoI	Quantidade de tentativas	
	Por palavras-chave	Por tags
Eiffel Tower	3	4
Garnier Opera	1	5
Sacré-Cœur Basilica	2	5
The Louvre	2	1
Pompidou Center	1	5
Montparnasse Tower	1	5
Les Invalides	4	3
Notre Dame de Paris	1	2
Place de la République	1	2
Total	16	32

quantidade necessária de consultas para a obtenção de informações dos PoIs identificados considerando ambos os cenários.

O experimento consistiu de uma busca por PoIs em Paris - França e utilizou os mesmos parâmetros dos experimentos anteriores (ver Tabela 6). Para cada PoI identificado, foram feitas tentativas de obtenção das informações utilizando tanto as palavras-chave quanto as tags com maiores pontuações. Como apresentado na tabela, $N_{pc} = 4$. Por essa razão, apenas as quatro primeiras tags e as quatro primeiras palavras-chave foram utilizadas nas buscas. Dessa forma, se, por exemplo, a terceira palavra-chave obteve sucesso na consulta à fonte de informações, o número de tentativas utilizadas é três. Se a busca por proximidade geográfica foi realizada para um determinado PoI, então cinco tentativas foram gastas (quatro utilizando as tags ou palavras-chave e uma com a busca por proximidade).

A Tabela 9 apresenta os locais identificados e que demandaram uma quantidade diferente de tentativas comparando-se os dois cenários. Os demais PoIs identificados não são apresentados, visto que, por demandarem a mesma quantidade de tentativas nos dois casos, não são úteis para a distinção dos cenários. Os resultados obtidos mostram que, de modo geral, os locais são melhor representados pelas palavras-chave, o que permitiu a obtenção das informações com metade das consultas ao repositório de informações.

5.4 Tempo de execução

A Tabela 10 apresenta os tempos de execução do método ToPI para duas consultas, em Paris e em Nova York, ambas realizadas conforme os valores de parâmetros da Tabela 6. A análise do tempo foi feita por meio da execução do sistema em um computador com 4GB de memória RAM e processador Core 2 Duo 2.20 GHz. É importante ressaltar que a implementação não foi otimizada visando melhor tempo de execução e que os

Tabela 10 – Tempo de execução de cada etapa do método ToPI (em segundos).

Etapa	Paris (25 PoIs)	Nova York (23 PoIs)
Obtenção das fotografias	8,485	7,369
Agrupamento e filtragem	2,151	1,823
Pontuação das <i>tags</i>	1,287	1,359
Identificação das palavras-chave	0,035	0,006
Remoção de PoIs redundantes	0,257	0,53
Obtenção das informações ⁴	0,867	0,803
Total	13,082	11,89

tempos demandados por algumas etapas do método dependem do tempo de resposta de serviços externos, como o tempo demandado pela obtenção das fotografias no repositório de imagens e pela obtenção das informações dos PoIs.

Como pode ser observado, o tempo gasto na identificação das palavras-chave é muito pequeno, enquanto a utilização delas, conforme apresentado na Seção anterior, permite uma economia de requisições e consequente redução no tempo de processamento da etapa de obtenção das informações.

5.5 Considerações Finais

Este capítulo apresentou os experimentos realizados para a avaliação do método proposto. Os testes feitos mostraram que houve bom grau de concordância entre os locais retornados pelo método e aqueles pertencentes ao TripAdvisor e Google Maps, mostrando que, de fato, os locais exibidos ao usuário são relevantes. Além disso, foram apresentadas situações que mostraram a capacidade natural de atuar em contexto dinâmico, permitindo identificar novos PoIs, PoIs que deixaram de ser interessantes de acordo com as fotografias e PoIs já existentes que, por qualquer motivo, destacaram-se recentemente. A importância da utilização de palavras-chave em detrimento da utilização direta das *tags* com maiores pontuações foi demonstrada em um experimento e o tempo de execução do método, embora parcialmente dependente de serviços externos, foi apresentado por meio de duas buscas por PoIs.

Como pôde ser visto, situações como “lugar vs evento”, “PoIs dentro de PoIs” e outras podem ocorrer. Elas não representam erro e podem ser melhor investigadas pelo usuário com os próprios recursos do sistema a fim de se evitar enganos.

⁴ Tempo médio por PoI calculado a partir da obtenção das informações de cinco PoIs.

Conclusão

Este trabalho apresentou uma abordagem que utiliza fotografias e metadados provenientes de repositórios *online* de imagens para a identificação, não-supervisionada e em tempo real, de pontos turísticos próximos a qualquer localização ao redor do mundo. O método proposto trata naturalmente a dinamicidade desse tipo de repositório, onde fotografias são inseridas e retiradas o tempo todo. Assim, ao contrário do que ocorre com abordagens que utilizam informações provenientes de bases estáticas, o método é capaz de identificar PoIs recentemente abertos (ou que, por qualquer motivo, destacaram-se recentemente) ao mesmo tempo em que deixa de retornar locais que já não existem mais (ou existem mas deixaram de ser interessantes por qualquer motivo). Além disso, por atuar em contexto dinâmico, ele é capaz de refletir acontecimentos recentes, que podem influenciar (positivamente ou não) a decisão de visitar determinados locais.

O método proposto foi positivamente avaliado, de forma que os locais por ele identificados estão de acordo com PoIs exibidos pelo TripAdvisor e/ou Google Maps nos cenários analisados. Algumas situações especiais eram esperadas e foram encontradas durante os experimentos. Elas não representam erros e podem ser melhor investigadas pelo usuário por meio de ferramentas de interação apresentadas neste trabalho caso gerem qualquer tipo de dúvida.

Por fim, cabe destacar os dois principais meios de utilização da abordagem proposta nesta dissertação:

1. Sistema *ToPI Trip*: O sistema apresentado oferece ao usuário ferramentas para a exploração do conhecimento obtido pelo método e permite que ele planeje e tome decisões sobre visitas aos pontos turísticos.
2. Saída do método como entrada em outras abordagens: Os PoIs identificados podem servir como dados de entrada em sistemas mais complexos, como sistemas de recomendação turística. Assim, a necessidade da utilização de base de dados estáticas para esse fim poderia ser reduzida ou evitada.

6.1 Limitações e Impactos

Alguns fatores exercem influência no método e determinam sua aplicabilidade e resultados. São eles:

- ❑ Quantidade de imagens e quantidade/relevância das *tags*. É preciso que haja uma quantidade mínima de imagens que permita a obtenção de ao menos um grupo (que, após a filtragem segundo a quantidade mínima de usuários, possa vir a ser considerado um PoI). Além disso, quanto mais e melhores as *tags*, maior probabilidade de que as palavras-chave representem fielmente os locais, favorecendo a busca por informações.
- ❑ Escolha dos valores dos parâmetros *Eps* e *MinPts* utilizados no agrupamento das fotos. A escolha de valores fixos e globais para esses parâmetros pode prejudicar a identificação de PoIs, tendo em vista que o tamanho de um PoI pode ser bastante diferente em relação ao tamanho de outro e que não se pode assumir que haja uma distribuição semelhante das imagens para quaisquer PoIs.
- ❑ Qualidade da fonte de informações sob o ponto de vista de:
 1. Quantidade de locais conhecidos em sua base de dados e frequência de atualização, a fim de possibilitar a obtenção de informações de PoIs (1) presentes nas mais diferentes localidades pesquisadas e (2) recentes.
 2. Capacidade de obter resultados em buscas cujas palavras-chave não sejam idênticas aos valores retornados;
 3. Tempo de resposta: Como o trabalho proposto atua em tempo real, é importante que a fonte de informações utilizada seja ágil em relação ao tempo de resposta, a fim de não prejudicar essa característica da proposta;
 4. Diferença de idiomas: A consulta feita na fonte de informações utiliza uma *tag* (palavra-chave) cujo idioma não pode ser previsto. Por isso, a capacidade da fonte em obter resultados independentemente do idioma utilizado na consulta impacta diretamente na capacidade de obtenção das informações.

6.2 Trabalhos Futuros

Apesar do bom resultado alcançado, alguns aspectos podem ser investigados a fim de se melhorar algumas etapas do processo ou adicionar novas capacidades nele:

- ❑ Agrupamento das imagens: Os valores dos parâmetros *Eps* e *MinPts* utilizados ao longo do trabalho podem variar de acordo com a localização geográfica considerada, tendo em vista a diferença de tamanhos entre PoIs e suas distribuições de imagens.

Dessa forma, meios de automatizar a escolha desses valores podem ser estudados e aplicados. Uma das estratégias da literatura para esse fim pode ser vista em (KARAMI; JOHANSSON, 2014).

- ❑ **Visualização dos resultados:** Uma abordagem bastante utilizada na visualização de locais turísticos é a exibição em mapas. Essa estratégia, entretanto, pode não ser intuitiva ou não atender o usuário caso ele queira, por exemplo, ver simultaneamente informações de dois PoIs distantes entre si geograficamente para realizar uma comparação entre ambos e decidir qual visitar. Por essa razão, abordagens visuais que promovam diferentes meios de exploração dos pontos de interesse identificados também podem ser investigadas.
- ❑ **Situação “Lugar vs evento”:** Alguns trabalhos, como (LUO et al., 2008), (YUAN et al., 2008) e (RATTENBURY; GOOD; NAAMAN, 2007), propõem métodos para a diferenciação de locais e eventos. Abordagens desse tipo podem ser estudadas e aplicadas ao método deste trabalho de forma a prover uma personalização de busca mais sofisticada.
- ❑ **Demais situações e remoção de PoIs redundantes:** Formas mais sofisticadas de tratamento para situações como “PoIs dentro de PoIs” e a redundância de PoIs podem ser estudadas e aplicadas a fim de melhorar a qualidade das informações retornadas.
- ❑ **Conteúdo visual:** Pode ser que a utilização do conteúdo visual das imagens melhore a qualidade do agrupamento e, conseqüentemente, do método. Um estudo detalhado precisa ser feito para analisar se a utilização desse tipo de informação poderá ser feita mantendo as características de tempo real e contexto dinâmico existentes atualmente.

Referências

AGRAWAL, R. et al. Automatic subspace clustering of high dimensional data for data mining applications. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 27, n. 2, p. 94–105, jun. 1998. ISSN 0163-5808. Disponível em: <<http://doi.acm.org/10.1145/276305.276314>>.

AHERN, S. et al. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: ACM. **Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries**. [S.l.], 2007. p. 1–10.

ANKERST, M. et al. Optics: Ordering points to identify the clustering structure. In: **Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99)**. Philadelphia, PA: [s.n.], 1999. p. 49–60.

AVRITHIS, Y. S. et al. Retrieving landmark and non-landmark images from community photo collections. In: BIMBO, A. D.; CHANG, S.-F.; SMEULDERS, A. W. M. (Ed.). **ACM Multimedia**. ACM, 2010. p. 153–162. ISBN 978-1-60558-933-6. Disponível em: <<http://dblp.uni-trier.de/db/conf/mm/mm2010.html#AvrithisKTS10>>.

_____. **Visual Image Retrieval and Localization Demo**. 2010. <<http://viral.image.ntua.gr/>>. Accessed: 2015-07-18.

BRILHANTE, I. et al. Where shall we go today?: planning touristic tours with tripbuilder. In: HE, Q. et al. (Ed.). **CIKM**. [S.l.]: ACM, 2013. p. 757–762. ISBN 978-1-4503-2263-8.

_____. Tripbuilder: A tool for recommending sightseeing tours. In: RIJKE, M. de et al. (Ed.). **Advances in Information Retrieval**. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8416). p. 771–774. ISBN 978-3-319-06027-9. Disponível em: <http://dx.doi.org/10.1007/978-3-319-06028-6_93>.

_____. **TripBuilder Demo**. 2014. <<http://tripbuilder.isti.cnr.it/>>. Accessed: 2015-02-10.

CAMPELLO, R.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 7819). p. 160–172. ISBN 978-3-642-37455-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-37456-2_14>.

CAMPELLO, R. J. G. B. et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 10, n. 1, p. 5:1–5:51, jul. 2015. ISSN 1556-4681. Disponível em: <<http://doi.acm.org/10.1145/2733381>>.

CAO, L. et al. A worldwide tourism recommendation system based on geotaggedweb photos. In: **ICASSP**. IEEE, 2010. p. 2274–2277. ISBN 978-1-4244-4296-6. Disponível em: <<http://dblp.uni-trier.de/db/conf/icassp/icassp2010.html#CaoLGJHH10>>.

CNET. **Google commutes death sentence for Panoramio photo site**. 2015. <<http://www.cnet.com/news/google-commutes-death-sentence-for-panoramio-photo-site/>>. Accessed: 2015-07-28.

COMANICIU, D.; MEER, P. Mean shift analysis and applications. In: IEEE. **Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on**. [S.l.], 1999. v. 2, p. 1197–1203.

CRANDALL, D. et al. Mapping the world's photos. In: **WWW**. [S.l.: s.n.], 2009.

DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: _____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. cap. Encyclopedia of Distances, p. 1–583. ISBN 978-3-642-00234-2. Disponível em: <http://dx.doi.org/10.1007/978-3-642-00234-2_1>.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

Flickr. **Serviços do Flickr: API: flickr.photos.search**. 2015. <<https://www.flickr.com/services/api/flickr.photos.search.html>>. Accessed: 2015-04-05.

GAGGI, O. Discovering local attractions from geo-tagged photos. In: **Proceedings of the 28th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2013. (SAC '13), p. 730–735. ISBN 978-1-4503-1656-9. Disponível em: <<http://doi.acm.org/10.1145/2480362.2480502>>.

GOWER, J. C.; ROSS, G. Minimum spanning trees and single linkage cluster analysis. **Applied statistics**, JSTOR, p. 54–64, 1969.

GUHA, S.; RASTOGI, R.; SHIM, K. Cure: An efficient clustering algorithm for large databases. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 27, n. 2, p. 73–84, jun. 1998. ISSN 0163-5808. Disponível em: <<http://doi.acm.org/10.1145/276305.276312>>.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques: concepts and techniques**. [S.l.]: Elsevier, 2011.

HE, W. et al. Generating landmark overviews with geo-tagged web photos. In: **SMC**. IEEE, 2012. p. 2644–2649. ISBN 978-1-4673-1713-9. Disponível em: <<http://dblp.uni-trier.de/db/conf/smc/smc2012.html#HeLWHL12>>.

HIJMAN, R. J. **Introduction to the "geosphere" package (Version 1.5-1)**. [S.l.], 2015.

Instagram. **Instagram Press**. 2015. <<https://www.instagram.com/press/>>. Accessed: 2015-07-28.

- _____. **Media Endpoints - Instagram Developer Documentation**. 2015. <<https://www.instagram.com/developer/endpoints/media/>>. Accessed: 2015-07-28.
- IVIS, F. Calculating geographic distance: Concepts and methods. In: **Proceedings of the 19th Conference of Northeast SAS User Group**. [S.l.: s.n.], 2006.
- JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recogn. Lett.**, Elsevier Science Inc., New York, NY, USA, v. 31, n. 8, p. 651–666, jun. 2010. ISSN 0167-8655. Disponível em: <<http://dx.doi.org/10.1016/j.patrec.2009.09.011>>.
- KARAMI, A.; JOHANSSON, R. Choosing dbSCAN parameters automatically using differential evolution. **International Journal of Computer Applications**, Foundation of Computer Science (FCS), v. 91, n. 7, p. 1–11, 2014.
- KAUFMAN, L.; ROUSSEEUW, P. Clustering by means of medoids. **Statistical Data Analysis Based on the L1-Norm and Related Methods**, p. 405–416, 1987.
- KISILEVICH, S.; MANSMANN, F.; KEIM, D. P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: **Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application**. New York, NY, USA: ACM, 2010. (COM.Geo '10), p. 38:1–38:4. ISBN 978-1-4503-0031-5. Disponível em: <<http://doi.acm.org/10.1145/1823854.1823897>>.
- LEMOIS, F. et al. Towards a context-aware photo recommender system. **Context-Aware Recommender System Workshops**, 2012.
- LUO, J. et al. Event recognition: Viewing the world with a third eye. In: **Proceedings of the 16th ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2008. (MM '08), p. 1071–1080. ISBN 978-1-60558-303-7. Disponível em: <<http://doi.acm.org/10.1145/1459359.1459574>>.
- MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). **Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.]: University of California Press, 1967. v. 1, p. 281–297.
- MAJID, A. et al. A system for mining interesting tourist locations and travel sequences from public geo-tagged photos. **Data Knowl. Eng.**, v. 95, p. 66–86, 2015. Disponível em: <<http://dblp.uni-trier.de/db/journals/dke/dke95.html#MajidCMHC15>>.
- MICHEL. **How many public photos are uploaded to Flickr every day, month, year?** 2016. <<https://www.flickr.com/photos/franckmichel/6855169886/>>. Accessed: 2016-01-05.
- MONMA, Y. **Algoritmo Rápido para Segmentação de Vídeos Utilizando Agrupamento de Clusters**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 2014.
- RATTENBURY, T.; GOOD, N.; NAAMAN, M. Towards automatic extraction of event and place semantics from flickr tags. In: **Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2007. (SIGIR '07), p. 103–110.

- ISBN 978-1-59593-597-7. Disponível em: <<http://doi.acm.org/10.1145/1277741.1277762>>.
- SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Inf. Process. Manage.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988. ISSN 0306-4573. Disponível em: <[http://dx.doi.org/10.1016/0306-4573\(88\)90021-0](http://dx.doi.org/10.1016/0306-4573(88)90021-0)>.
- SHI, Y. et al. Nontrivial landmark recommendation using geotagged photos. **ACM TIST**, v. 4, n. 3, p. 47, 2013. Disponível em: <<http://dblp.uni-trier.de/db/journals/tist/tist4.html#ShiSHL13>>.
- SINNOTT, R. W. Virtues of the Haversine. **Sky Telesc.**, v. 68, p. 159, 1984.
- STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. In: **6th ACM SIGKDD, World Text Mining Conference**. [S.l.: s.n.], 2000.
- UNWTO. **UNWTO tourism highlights**. [S.l.]: United Nations World Tourism Organization Madrid, 2014.
- _____. **Over 1.1 billion tourists travelled abroad in 2014**. 2015. <<http://media.unwto.org/press-release/2015-01-27/over-11-billion-tourists-travelled-abroad-2014>>. Accessed: 2015-02-10.
- UOL. **O que já se sabe sobre os ataques terroristas em Paris**. 2015. <<http://noticias.uol.com.br/internacional/ultimas-noticias/2015/11/14/o-que-se-sabe-sobre-os-ataques-terroristas-em-paris.htm>>. Accessed: 2016-03-08.
- VINCENTY, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. **Survey Review**, v. 22, n. 176, p. 88–93, 1975.
- WANG, W.; YANG, J.; MUNTZ, R. R. Sting: A statistical information grid approach to spatial data mining. In: **Proceedings of the 23rd International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (VLDB '97), p. 186–195. ISBN 1-55860-470-7. Disponível em: <<http://dl.acm.org/citation.cfm?id=645923.758369>>.
- WEBB, A. **Statistical Pattern Recognition**. Arnold, 1999. (A Hodder Arnold Publication). ISBN 9780340741641. Disponível em: <<https://books.google.com.br/books?id=YCMicH3-JGQC>>.
- YUAN, J. et al. Mining gps traces and visual words for event classification. In: **Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval**. New York, NY, USA: ACM, 2008. (MIR '08), p. 2–9. ISBN 978-1-60558-312-9. Disponível em: <<http://doi.acm.org/10.1145/1460096.1460099>>.
- ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. Birch: An efficient data clustering method for very large databases. In: **Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 1996. (SIGMOD '96), p. 103–114. ISBN 0-89791-794-4. Disponível em: <<http://doi.acm.org/10.1145/233269.233324>>.