TripTag: Ferramenta de planejamento de viagens baseada em experiências de usuários de redes sociais

Antônio H. G. Leite, Fabrício Benevenuto, Mirella M. Moro[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais, Belo Horizonte, MG

{antonioh, fabricio, mirella}@dcc.ufmg.br

Abstract. *The new Internet-connected generations of tourists search for destinations and plan their trips through online services. A new way of doing so is through "travel social networks", in which regular people share their opinion about touristic places. In such networks, the information is always updated and not biased according to agencies interests. However, there are two major problems: searching for a place in such networks the user to know where she or he wants to go, and acquiring knowledge about it requires reading its reviews, which vary from a handful to hundreds of texts. In this paper, we describe a new, exciting tool that automatically crawls different travel social networks, aggregates their content, tags them, allows to search for such tags, performs sentiment analysis and summarizes the reviews through word clouds.*

Resumo. *As novas gerações de turistas buscam por destinos e planejam suas viagens através de serviços online. Uma das mais novas e populares maneiras de explorar tais serviços é através das recém criadas "redes sociais de viagens", nas quais usuários compartilham suas opiniões sobre pontos turísticos. A informação é sempre atualizada, e as opiniões não tendem de acordo com os interesses de agências. Porém, existem dois problemas: buscar por um local requer que o usuário saiba para onde quer ir, e adquirir conhecimento sobre ele requer a leitura de suas avaliações, o que pode variar desde um pequeno número até centenas de textos. Neste artigo, descrevemos uma ferramenta nova que automaticamente coleta dados de diferentes redes sociais de viagem, agrega seus conteúdos, adiciona tags, permite a busca por tags, realiza análise de sentimento e resume as avaliações em nuvem de palavras.*

1. Introdução

As novas gerações de turistas, que vivem em um mundo conectado, planejam suas viagens e buscam por destinos através de serviços online. Uma maneira nova e que tem se tornado popular é explorar tais serviços através de *travel social networks*, nas quais usuários compartilham suas opiniões sobre pontos turísticos famosos, bem como sobre lugares desconhecidos. Uma vantagem é que a informação é sempre nova e atualizada, diferente das tradicionais agências de viagens e publicações específicas. Outra é a possibilidade de identificar o que várias pessoas estão pensando e qual o sentimento delas sobre o lugar, ao invés de apenas ter acesso à opinião do autor de um livro ou dos agentes de viagem.

Entretanto, existem dois grandes problemas: buscar por um local em tais redes requer que o usuário saiba para onde quer ir, e adquirir conhecimento sobre ele requer a

leitura de suas avaliações, o que pode variar de um pequeno número a centenas de textos. Em outras palavras, é possível buscar pela cidade de Nova Iorque, mas não por uma "cidade grande e cosmopolita". Uma vez encontrada a cidade, o usuário precisa ler vários textos antes de formar sua opinião. Desse modo, apesar das redes sociais de viagem serem um grande avanço, elas ainda estão longe de facilitar a vida do usuário, que precisa de muito tempo de leitura para entender sobre os possíveis destinos.

O objetivo deste artigo é demonstrar as etapas de criação e funcionamento de uma ferramenta, chamada *TripTag*, para auxiliar na busca por pontos turísticos. As suas principais contribuições são assim resumidas: (*i*) a *TripTag* reune informações turísticas a partir das *várias* redes sociais existentes em um banco de dados; (*ii*) a partir dos dados, são geradas anotações (tags) que resumem as avaliações (escritas pelos usuários das redes) de cada local; (*iii*) com base nessas anotações, a ferramenta gera uma nuvem de termos mais frequentes e a análise de sentimento das avaliações; e (*iv*) a busca por locais pode ser agora realizada através das tags. É importante notar que a nuvem de termos e a análise de sentimento são consideradas excelentes resumos sobre as avaliações dos locais, facilitando assim a vida do usuário durante sua busca. Além da fácil visualização, a busca também é aperfeiçoada através das tags, com as quais o usuário pode buscar pelas características dos locais que deseja visitar.

A seguir, apresentamos uma breve descrição de esforços relacionados, a arquitetura da ferramenta, sua interface e funcionamento, e, finalmente, apresentamos conclusões e direções para trabalhos futuros.

2. Trabalhos Relacionados

As redes sociais podem ser usadas como valiosa fonte de dados sobre perfis e preferências de usuários devido ao grande volume e tipo de informação compartilhada. Nesse contexto, várias aplicações surgiram na tentativa de explorar informações postadas em redes sociais, incluindo a análise de campanhas políticas [Tumasjan et al. 2010], a repercusão de variações em bolsas de valores [Bollen et al. 2010], a detecção e síntese da repercussão de desastres naturais [Sakaki et al. 2010] e epidemias [Gomide et al. 2011].

Entretanto, nenhum dos trabalhos citados aborda a construção de um sistema que, de forma automática, coleta avaliações de várias redes sociais relacionadas a viagens, gera tags e possibilita que sejam feitas buscas baseadas nessas tags. Geralmente, os sistemas existentes permitem que a busca seja realizada pelo nome da cidade ou região desejada e em apenas uma rede social. As tags geradas são utilizadas para mostrar ao usuário uma visão resumida (em nuvem de palavras) sobre o que as pessoas estão escrevendo sobre o local e qual o sentimento delas em relação aos locais. Sendo assim, nosso aplicação é inovadora e complementar aos esforços existentes na literatura.

3. Arquitetura da Ferramenta

O funcionamento da ferramenta é dividido nas seguintes etapas: obtenção dos dados, mineração de dados e análise de sentimentos. A Figura 1 ilustra tais etapas e o fluxo de dados entre elas. Cada etapa é detalhada a seguir.

3.1. Coleta de Dados

O princípio fundamental para o desenvolvimento da ferramenta é utilizar apenas informações turísticas retiradas de redes sociais. Dessa forma, o primeiro passo foi

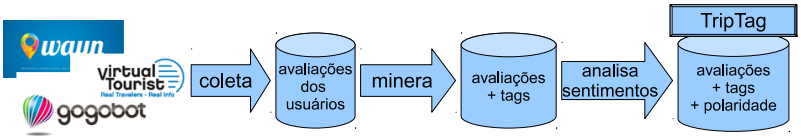

Figura 1. Fluxo de dados e processos

escolher as fontes de dados mais adequadas dentre as várias redes sociais existentes. Inicialmente, foram escolhidas três redes sociais de viagem que são populares e possuem dados que facilitam a extração das informações dadas pelos usuários. São elas: *Where are you now?* (WAYN)[1] - a maior com 21 milhões de usuários, Gogobot[2] e Virtual Tourist[3].

Nessas redes sociais, os usuários se cadastram e possuem perfis com fotos e outras informações. As cidades também possuem perfis que são alimentados exclusivamente com informações que os próprios usuários postam através de suas contas. Dessa forma, todas as cidades nas redes sociais possuem uma página com uma enorme quantidade de informações provindas de pessoas que já as visitaram. Tal característica é a principal justificativa para termos escolhidos esses sites.

Porém, tais redes sociais não possuem APIs para a extração de dados. Desse modo, a primeira etapa é **coletar** os dados de tais redes através de *crawlers* específicos para cada website. Para viabilizar uma coleta mais rápida, foi definido um subconjunto de cidades para serem as sementes da coleta. O critério utilizado para a escolha das cidades foi a seleção das 200 cidades com o maior número de visitantes no mundo, pois essas seriam as com maior número de avaliações, provendo assim um bom volume de dados a serem processados. As coordenadas geográficas de cada cidade foram obtidas através da API do Google Maps e armazenadas no banco de dados. Em resumo, foram desenvolvidos três *crawlers* (um por website) que visitam o perfil das 200 cidades na respectiva rede social, identificam as avaliações dos usuários dessa cidade e as salvam em um banco de dados relacional.

3.2. Mineração de Dados

A segunda etapa é a mineração dos dados obtidos na etapa anterior. Tal etapa visa **gerar anotações** (tags) para cada cidade. Para cada cidade, são identificadas as palavras-chave e expressões que aparecem mais vezes nos textos e no maior de número de textos. A hipótese é: se uma palavra ou expressão é muito utilizada nas avaliações dos usuários sobre uma cidade, ela se relaciona fortemente com as experiências vividas pelo turista quando a visitou. As palavras ou expressões mais frequentes definem as tags para cada cidade. No contexto deste trabalho, não foram exploradas as relações entre os usuários das redes sociais escolhidas, o que será feito em trabalhos futuros. Nosso intuito foi utilizar as revisões dos usuários como uma fonte *crowsourced* de informações turísticas.

O algoritmo utilizado é uma modificação do Apriori [Zaki and Meira Jr. 2013]. Inicialmente, o algoritmo faz a contagem das expressões com uma palavra e elimina todas as que forem menos frequentes que um determinado limiar. Em seguida, é calculada a frequência das expressões com duas palavras, e novamente são excluídas expressões

[1]WAYN: `http://www.wayn.com`
[2]Gogobot: `http://www.gogobot.com`
[3]Virtual Tourist: `http://www.virtualtourist.com`

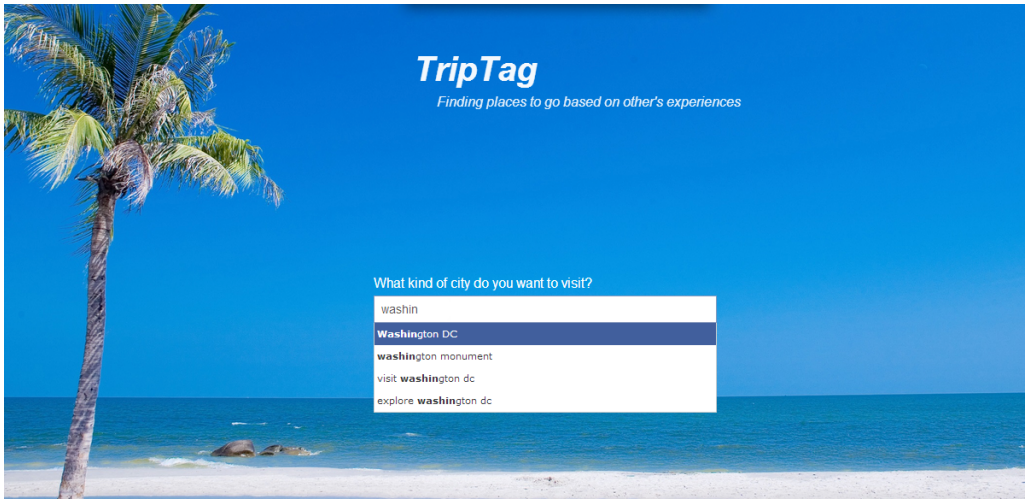

Figura 2. Interface inicial da ferramenta

infrequentes. O algoritmo segue elevando o grau desse *n-grama* a cada ciclo, e faz cortes até não ter mais expressões frequentes. Duas observações importantes: só são considerados os n-gramas formados por palavras adjacentes, e o processo também se encarrega de eliminar *stopwords* como artigos e conjunções.

3.3. Análise de Sentimentos

Finalmente, é realizada a **análise de sentimento** das avaliações. Existem várias formas de se realizar análise de sentimento, que variam desde abordagens léxicas até abordagens que utilizam técnicas de aprendizagem de máquina [Gonçalves et al. 2013]. Neste trabalho, por simplicidade, utilizamos uma biblioteca que implementa um classificador Naive Bayes que se encontra disponível em `http://text-processing.com/docs/sentiment.html`. Em trabalhos futuros, pretendemos investigar outras abordagens e avaliar qual a mais adequada para o nosso contexto.

Cada avaliação textual é submetida ao classificador de sentimentos que retorna a probabilidade com a qual o sentimento associado ao texto é positivo, neutro ou negativo. Esses três valores associados a cada avaliação são então armazenados no banco de dados.

4. Interface e Funcionamento

Após o banco de dados ser populado com as avaliações e suas tags, a ferramenta TripTag processa as consultas do usuário sobre tais dados. O sistema é baseado na linguagem PHP e no framework CakePHP[4], que utilizada o modelo de desenvolvimento de software *Model-view-controller* (MVC).

A ferramenta foi projetada para ser simples e fornecer visualizações para facilitar a compreensão das avaliações. Consequentemente, ela permite identificar facilmente o que os visitantes pensam sobre cada cidade. A tela inicial é composta apenas por uma barra de busca na qual é possível fazer dois tipos de busca: pelo nome da cidade ou por tags. Ao digitar qualquer texto, o sistema busca no banco de dados nomes de cidades ou tags que contenham a palavra digitada. Os nomes de cidades ou tags encontradas são exibidos

[4]CakePHP: `http://cakephp.org/`

Figura 3. Resultado para a cidade de Las Vegas

em uma lista, conforme ilustrado na Figura 2. Ao escolher uma palavra ou expressão específica, o usuário é redirecionado para o perfil da cidade. Se o usuário optar por uma tag, é criado um token na barra de buscas e novas tags podem ser escolhidas conforme o desejado. Depois de escolhidas as tags, ao clicar no botão <Search>, é exibida uma lista de cidades que contêm todas as tags utilizadas na busca e, ao clicar no nome de uma cidade, o usuário é direcionado para o perfil dela.

O resultado para uma busca que identificou a cidade de Las Vegas é ilustrado na Figura 3. É importante notar que o perfil de cada cidade também exibe um mapa com a sua localização no planisfério, uma nuvem de termos com as tags associadas a essa cidade e um gráfico que sumariza as análises de sentimento de todas as avaliações daquela cidade. Além disso, é exibida uma listagem de todas as avaliações na qual é possível que sejam feitas buscas através do campo <search>.

Outro exemplo é ilustrado na Figura 4, com as informações sobre a cidade de Chennai (anteriormente conhecida como Madras, na Índia). Diferentemente de Las Vegas, Chennai é conhecida pelos seus templos e possui avaliações mais positivas: acima de 60%, ao contrário de Las Vegas que ficou abaixo de 45%.

5. Conclusão

Este artigo apresentou a *TripTag*, uma ferramenta para busca de destinos de viagem que funciona em um banco de dados criado a partir de avaliações turísticas provenientes de redes sociais de viagens. A partir da contagem da frequência das palavras nas avaliações, são definidas tags relevantes baseadas apenas nas experiências relatadas pelos usuários das redes sociais. Essas tags serão agrupadas em uma nuvem de termos. As avaliações também passam pelo processo de análise de sentimentos a fim de sumarizar as incontáveis opiniões sobre cada cidade.

Observamos que existe um grande número de spams e avaliações com erros ortográficos, os quais dificultam a análise dos dados. Desse modo, como trabalho futuro,

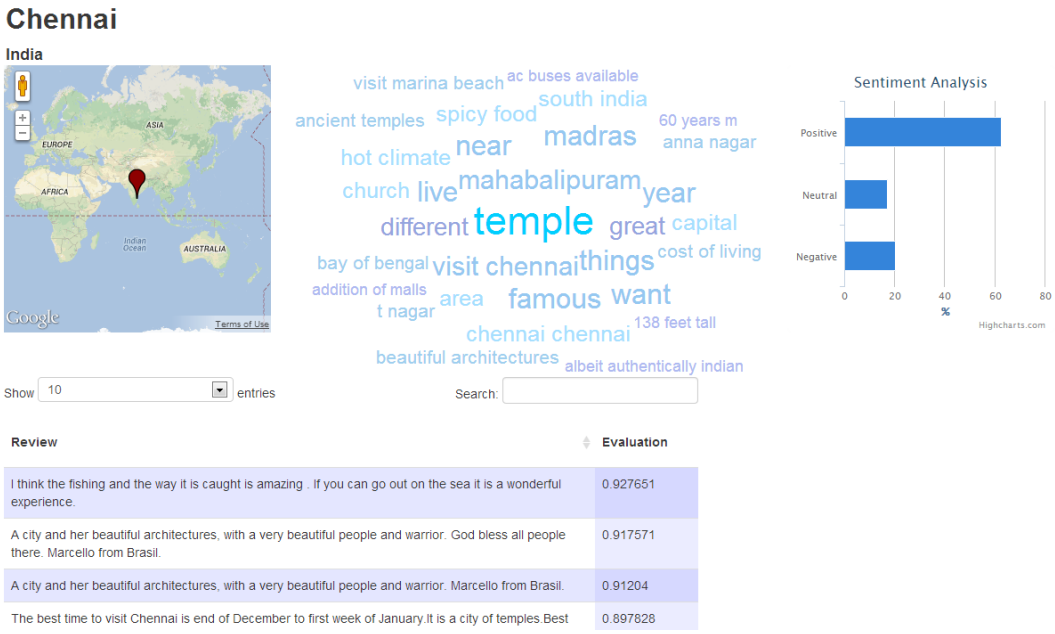

Figura 4. Resultado para a exótica Chennai

existe a possibilidade de melhorar os resultados das tags e da análise de sentimentos através da utilização de técnicas de *data cleaning*, por exemplo. Além disso, serão estudadas as relações existentes entre os usuários das redes sociais de viagem e como elas podem ser exploradas para o melhoramento da ferramenta. Também será desenvolvida uma forma de avaliar os resultados obtidos através da ferramenta. Finalmente, um screencast da ferramenta está disponível em:

http://www.dcc.ufmg.br/∼mirella/TripTag

Agradecimentos: Esse trabalho foi financiado por CNPq, FAPEMIG e InWeb, Brasil.

Referências

[Bollen et al. 2010] Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

[Gomide et al. 2011] Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Procs. of ACM WebSci*.

[Gonçalves et al. 2013] Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Procs. of ACM COSN*.

[Sakaki et al. 2010] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Procs. of WWW*.

[Tumasjan et al. 2010] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Procs. of ICWSM*.

[Zaki and Meira Jr. 2013] Zaki, M. and Meira Jr., W. (2013). *Data Mining and Analysis: Foundations and Algorithms*. Cambridge Un. Press (Draft available at `http://bit.ly/12WMuem`).