Análises de Dados de Sistemas Crowdsourcing: estudo de caso de avaliações de estabelecimentos realizadas no Yelp

Mateus P. Silveira[1], **Wender Z. Xavier**[1], **Humberto T. Marques-Neto**[1]

[1]Programa de Pós-Graduação em Informática
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC-MG)
Belo Horizonte – MG – Brasil

{mateus.parreiras, wender.xavier}@sga.pucminas.br, humberto@pucminas.br

Abstract. *This work does an analysis of the Yelp database, a commercial appraisal platform, which is popular in Europe and the United States. A characterization and analysis of feelings were made to determine the behavior of the users of this platform to help not only the improvement of the services provided by the establishments but also to contribute with a better understanding of the dynamics of the use of the services in a city.*

Resumo. *Este trabalho faz uma análise da base de dados do Yelp, uma plataforma para avaliação de estabelecimentos comerciais, muito popular na Europa e EUA. Realizou-se uma caracterização e uma análise de sentimentos para determinar o comportamento dos usuários desta plataforma para auxiliar não somente o aprimoramento dos serviços prestados pelos estabelecimentos como também contribuir com um melhor entendimento da dinâmica de utilização dos serviços em uma cidade.*

1. Introdução

A compreensão do comportamento humano acerca da utilização dos serviços da cidade é um tema amplamente pesquisado para melhoria de serviços e no desenvolvimento de cidades inteligentes [Batty et al. 2012]. O desenvolvimento dessas aplicações podem ainda, a partir do comportamento humano, avaliar e indicar melhorias no cotidiano das pessoas de forma a evitar riscos de acidentes e hábitos prejudiciais à saúde [Gustafson et al. 2014].

Aplicações de *Crowdsourcing* possuem um papel muito importante na coleta e na distribuição de informações. Estas aplicações são responsáveis por distribuir uma ou mais tarefas para que uma comunidade de pessoas, podendo ser questionários, tirar fotos e fazer resenhas (e.g Wikipédia[1] - Enciclopédia escrita de maneira colaborativa, ReclameAqui[2] - Site de reclamações contra empresas sobre atendimento, compra, serviços, etc.).

Com objetivo de analisar padrões de comportamento humano e verificar características de bases de dados de aplicações *Crowdsourcing*, utilizamos neste trabalho dados disponibilizados pela plataforma *Yelp*. Esta plataforma conta com um site e um aplicativo de avaliação de estabelecimentos comerciais.Em 2014 a *Yelp* disponibilizou parte de sua base de dados para que a academia pudesse utilizar esses dados em pesquisas. Então,

[1]http://www.wikipedia.org
[2]http://www.reclameaqui.com.br

uma análise da base de dados foi feita com o intuito de melhor entender o conjunto de dados. Assim, foi possível encontrar certos padrões como o grande número de avaliações 5 estrelas, o crescimento das de 1 estrela nos últimos anos e que normalmente os usuários tendem a fazer somente uma avaliação ao invés de várias. Além disso, uma análise de sentimento foi empregada para determinar se existia diferenças em textos com quantidade de estrelas dadas, que demonstrou que normalmente as avaliações se mantêm neutras.

Este trabalho está organizado da seguinte maneira. Seção 2 apresenta alguns trabalhos relacionados à aplicações de *Crowdsoucing*. Seção 3 apresenta a base de dados Yelp, principais informações contidas e propostas de análise. Seção 4 apresenta resultados obtidos a partir da análise da base do Yelp. A conclusão e trabalhos futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

O trabalho [Zhang 2015] enfatiza a importância de incorporar revisões textuais para recomendação através de análise de sentimento de nível de frase e investigar ainda mais o papel que os textos desempenham em diversas tarefas importantes de recomendação. Em [Yu et al. 2014], os autores prepõem combinar informações de relacionamento heterogêneas para cada usuário de forma diferente e fornecer resultados de recomendação personalizados de alta qualidade usando dados de feedback implícitos do usuário e modelos de recomendação personalizados.

[McClanahan and Gokhale 2016] propõem uma nova abordagem para entender as relações entre os clientes e as empresas e o tipo de informação que pode ser inferida a partir dessas relações. Um grafo é gerado com os nós sendo os negócios e o peso das arestas a quantidade de clientes em comum, e concluiu-se que os clientes preferem visitar empresas que são geograficamente próximos e/ou possuem produtos e serviços similares. Em [Bhowmick et al. 2017] os autores definem métricas de popularidade para rotular diversas empresas no conjunto de dados do *Yelp*, a principal foi a difusão da informação. Com isso, foi desenvolvido um modelo de recomendação que sugere as principais regiões para os empresários para iniciar negócios populares.

Os trabalhos apresentados tem como objetivo criar modelos de recomendação para o usuário, levando em consideração vários fatores. Nenhum dos trabalhos fazem uma comparação entre a análise de sentimento das avaliações e a quantidade de estrelas dadas comparando a base toda. Além disso, este trabalho faz uma caracterização da base de dados utilizada para que seja possível ter um melhor entendimento de maneia geral de como é o comportamento dos usuários.

3. Metodologia

Nesta Seção, as etapas da metodologia aplicadas no trabalho são apresentadas. Na primeira Subseção é apresentada a base de dados do *Yelp*. Em seguida é mostrado principais características e os processos de análise empregado nas avaliações.

3.1. Base de Dados *Yelp*

Em 2014 a empresa *Yelp* iniciou o desafio *Yelp Dataset Challenge*[3], disponibilizando uma base de dados contendo um subconjunto dos estabelecimentos comerciais, as avaliações e

[3]https://www.yelp.com/dataset/challenge

os usuários que utilizaram o aplicativo. O desafio já teve várias rodadas, na qual a base de dados era atualizada. A base de dados possui 5.200.000 avaliações, 174.000 negócios de 11 regiões metropolitanas de 4 países e 1.300.000 usuários, com dados de 2004 a 2017.

3.2. Caracterização da Base

Uma análise inicial na base de dados foi feita com intuito de compreender o seu conteúdo. Dentre as 20 cidades com maior número de estabelecimentos na base de dados, Las Vegas é a cidade que possui o maior conjunto de dados possuindo cerca de 25 mil estabelecimentos e 1,6 milhões de avaliações, seguida de Phoenix e Toronto. Para compreender o comportamento de usuário do *Yelp*, foi verificado primeiramente o número de avaliações por usuário. Destas 52,7% dos usuários do *Yelp* postam somente uma avaliação, 16,6% postaram duas avaliações, e 8,49% postaram três avaliações e somente 0,043% dos usuários fizeram mais de 250 avaliações, cerca de somente 500 no total. Assim, grande parte dos usuários tende utilizar a plataforma para conhecer informações e saber da experiência de outros usuários do que propriamente gerar conteúdo para a rede.

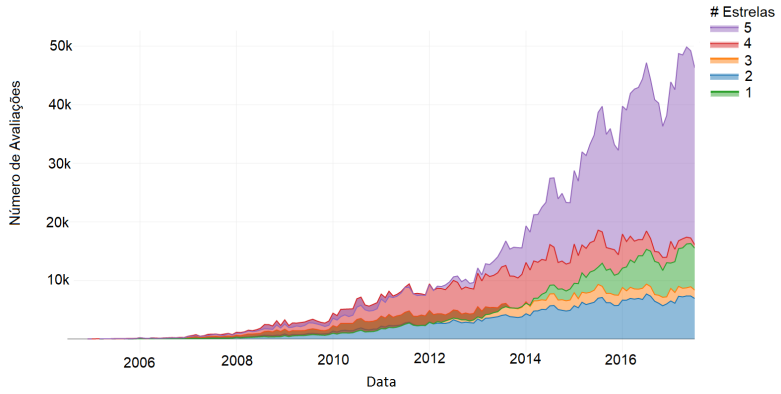


Figura 1. Número de avaliações por mês por estrela.

Os usuários podem dar notas de 1 a 5 estrelas para os estabelecimentos, sendo 1 experiencia ruim e 5 ótima. Na Figura 1 é possível perceber picos anuais entre os meses de Julho e Agosto, e o menor número de avaliações entre os meses de Novembro e Dezembro. Isso pode estar associado ao fato de os meses de Julho à Agosto serem períodos de férias escolares nos EUA, e os meses de Novembro e Dezembro as festas de fim-de-ano. Outro comportamento que pode ser visualizado é o número de estrelas atribuídas ao estabelecimento. A partir de 2014 existe uma tendência dos usuários darem mais avaliações de 1 estrela, enquanto notas de 2 a quatro tiveram pouca alteração. Com isso, pode-se perceber que os usuários estão tendendo a postar notas de 1 ou 5 estrelas para estabelecimentos que gostaram ou não gostaram.

Cada avaliação possui atributos como *usefull*, *cool* e *funny* que representam o número de usuários que acharam que a avaliação foi útil, interessante e engraçada respectivamente. Na Figura 2 é possível observar a correlação dentre estes atributos e o tamanho da avaliação (*text_length*). Uma correlação positiva muito forte existe entre os atributos *funny* e *usefull*, de 0,98, desta forma avaliações engraçadas costumam ser avaliações consideradas como úteis. Existe outras fortes correlações uma de 0,86 entre o tamanho do texto e o atributo *usefull*, e 0,76 entre *text_length* e *funny* demonstrando que o tamanho

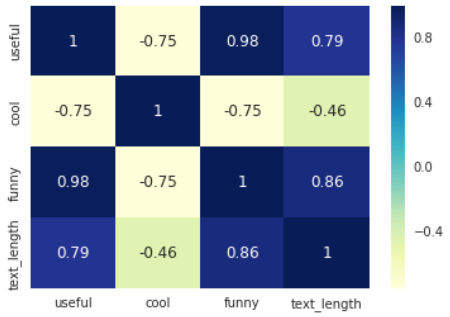


Figura 2. Correlação entre os atributos dos usuários.

do texto é importante para uma avaliação ser considerada como útil ou engraçada. Já a relação de *cool* e *usefull* é negativa, ou seja, se um aumenta o outro tende a diminuir.

A base de dados foi filtrada por selecionado as categorias de bar, restaurantes e comida de estabelecimentos nos EUA. Chegou-se à um conjunto de 49.914 estabelecimentos comerciais, 166.196 usuários e 569.291 avaliações. Então foi feita uma análise textual das avaliações para definir: quais são as palavras mais utilizadas por usuários dividida pela nota dada. Isto foi feito para verificar se avaliações com notas inferiores são mais negativas se comparadas com as avaliações com as notas mais altas. Para efetuar essas análises, foram utilizadas as bibliotecas do Python *NLTK*[4] *(Natural Language Toolkit)* e *TextBlob*[5]. Também foi feita a análise para definir se o sexo do usuário pode influenciar na maneira como ele posta comentários, utilizando a biblioteca *Gender-Guesser*[6]. Finalmente, para a criação do mapa de palavras foi utilizado o pacote Wordcloud[7] do R. Os resultados do processamento textual são apresentados na próxima Seção.

4. Resultados

Para efetuar a análise textual, cada avaliação passou por um processo de tratamento dos dados. Primeiramente foram retirados pontuação e em seguida todo o texto da avaliação foi transformado para letra minúscula. Em seguida, foram retiradas palavras de parada (i.e. *Stopwords*). Finalmente, o texto foi submetido à um processo de reduzir palavras flexionadas ao seu tronco (i.e. *stemming*). Então um mapa de palavras foi gerado para as avaliações considerando o número de estrelas (Figura 3) e três palavras se destacam em todos, *Food* (comida), *Place* (local), *Service* (serviço). Dito isto, percebeu-se que os três principais pontos que devem ser levados em conta para um estabelecimento de alimentação são a qualidade da comida, ambiente, e serviço prestado ao cliente, já que estes são as palavras mais escritas nas avaliações.

A Figura 4 mostra os resultados da aplicação dos algoritmos *TextBlob* e *NLTK* na base de dados. Ao analisar os resultados da aplicação do *TextBLob* (ver Figura 4(a)) percebe-se que em todas as categorias de estrelas o algoritmo encontrou valores variando entre -1 e 1, demonstrando que o número de estrelas de certa forma não está relacionado ao sentimento da avaliação. Entretanto, percebe-se que com o aumento do número de

[4]https://www.nltk.org/
[5]http://textblob.readthedocs.io/en/dev/
[6]https://pypi.python.org/pypi/gender-guesser/
[7]https://cran.r-project.org/package=wordcloud

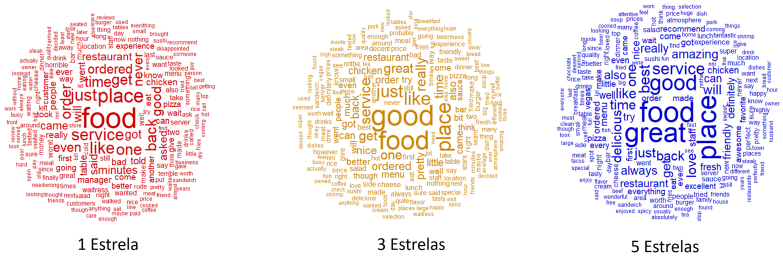
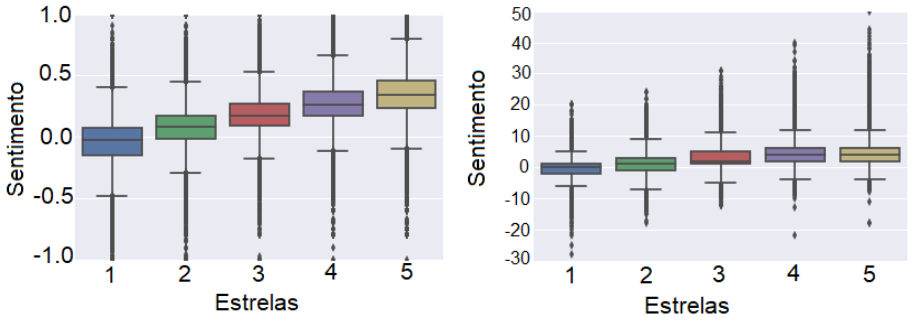


Figura 3. Mapa de Palavras das avaliações divididos pelo número de estrelas

estrelas atribuídas, cresce sutilmente o sentimento relacionado à avaliação. Grande parte das avaliações concentra-se em valores próximos à 0 para as avaliações com estrelas 1 e 2 e próximos à 0.5 para as avaliações de 5 estrelas. O mesmo comportamento pode ser observado nos resultados da análise de sentimento pelo *NLTK*, onde grande parte das avaliações se concentra próximo à valores de sentimentos neutros. Existem entretanto avaliações que chegaram à valores negativos como -20, e valores positivos próximos à 50 demonstrando que mesmo em avaliações boas dos estabelecimentos, os usuários podem estar utilizando de palavras consideradas negativas pelos algoritmos.



(a) Análise de sentimento utilizando o TextBlob.

(b) Análise de sentimento utilizando o NLTK.

Figura 4. Análise de sentimento das avaliações divididos pelo número de estrelas.

Uma última análise utilizando a biblioteca *Gender-Guesser* foi feita para verificar se o gênero das pessoas afeta na quantidade de estrelas das avaliações dadas. Esta biblioteca contém uma base de dados de nomes, e retorna uma estimativa do gênero da pessoa entre masculino, feminino e indefinido. Após a execução dos algoritmos, foi retornado 217.428 nomes masculinos, e 244.900 femininos. Considerando este valor, pôde-se perceber que ambos os gêneros utilizam a aplicação de forma parecida. Avaliando as estrelas atribuídas a cada avaliação este padrão se mantém, tendo o grupo feminino atribuído mais estrelas à uma quantidade maior de estabelecimentos comparado ao grupo masculino.

5. Conclusões e Trabalhos Futuros

Neste trabalho realizamos uma análise de base de dados de aplicações *Crowdsourcing* utilizando base de dados disponibilizada pelo *Yelp*. Identificamos os principais tópicos abordados na elaboração das avaliações dos clientes utilizando análise textual. Além

disso, Identificamos padrões de postagens ao longo do tempo que poderiam ser utilizados pelos estabelecimentos para realizar promoções para atrair maior número de clientes.

Além disso, foi apresentada uma análise inicial relacionado ao comportamento dos usuários. Verificamos que poucos usuários tendem a postar várias avaliações de estabelecimentos de diversas categorias, enquanto grande parte dos usuários tende postar somente uma avaliação. A cada avaliação pode ser definidos atributos em relação à humor, interesse e utilidade. Utilizando-se de bibliotecas do Python realizamos análise de sentimentos das avaliações esperando encontrar padrões entre a quantidade de estrelas e o sentimento do texto. Descobrimos que independente do número de estrelas atribuídas, as avaliações tendem a ter sentimento próximo à 0 (i.e. neutro).

Como trabalhos futuros existe a possibilidade de verificar séries temporais das avaliações e das estrelas atribuídas, verificando quando um restaurante pode prosperar ou não. Por fim, classificar os restaurantes por tipos mais específicos podem mostrar comportamentos diferentes de usuários para estabelecimentos diferentes. Novas modelagens de redes também representam um desafio na compreensão do comportamento de usuários do *Yelp* devido à complexidade e número de estabelecimentos e usuários na rede.

6. Agradecimentos

Este trabalho foi financiado por MASWeb (processo FAPEMIG/PRONEX APQ-01400-14), FAPEMIG (processo APQ-02924-16), PUC-Minas, CNPq, CAPES e STIC AmSud 18-STIC-07.

Referências

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., and Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518.

Bhowmick, A. K., Suman, S., and Mitra, B. (2017). Effect of information propagation on business popularity: A case study on yelp. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 11–20.

Gustafson, D. H., McTavish, F. M., Chih, M.-Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., Levy, M. S., Driscoll, H., Chisholm, S. M., Dillenburg, L., et al. (2014). A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry*, 71(5):566–572.

McClanahan, B. and Gokhale, S. S. (2016). Centrality and cluster analysis of yelp mutual customer business graph. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 592–601.

Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., and Han, J. (2014). Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 283–292, New York, NY, USA. ACM.

Zhang, Y. (2015). Incorporating phrase-level sentiment analysis on textual reviews for personalized recommendation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 435–440, New York, NY, USA. ACM.