Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Predição de links em uma rede heterogênea baseada em dados geolocalizados

Trabalho de Graduação

Camila Souto da Cunha Brendel Braga

Recife

Dezembro de 2017

Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Predição de links em uma rede heterogênea baseada em dados geolocalizados

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aluna: Camila Souto da Cunha Brendel Braga

Orientador: Ricardo Bastos Cavalcante Prudêncio

Recife, dezembro de 2017

"Forecasting is the art of saying what will happen, and then explaining why it didn't"
Anônimo

Agradecimentos

Gostaria, primeiramente, de agradecer a todos que fazem parte do Centro de Informática da Universidade Federal de Pernambuco, professores, funcionários e alunos.

Agradeço ao professor Ricardo Prudêncio, pela orientação neste trabalho e pelas discussões de ideias e problemas que encontrei durante o desenvolvimento.

Agradeço a In Loco, por permitir o uso do conjunto de dados nos experimentos deste trabalho, e pela oportunidade de fazer parte de algo incrível.

Agradeço à Maratona de Programação, assim como às professoras Liliane Salgado e Kátia Guimarães por guiarem o projeto. Agradeço também a todos que estavam comigo durante todos os sábados, mostrando que, se nos esforçamos o suficiente, conseguimos atingir melhores resultados que antes.

Agradeço ao PET-Informática, o melhor PET em linha reta do Brasil! Durante mais de três anos, o PET foi a melhor parte da minha graduação. Agradeço também ao professor Fernando Fonseca, FDFD, por guiar esse projeto durante anos.

Agradeço ao Clique: Maria Júlia Godoy (Júlia), Victor Monteiro (Vddm) e Simone Cohen (Simon). Com vocês passei pelos momentos mais engraçados e estressantes desses últimos cinco anos. Vocês marcaram a minha graduação e viraram amigos para a vida. Obrigada por tudo! Clique unido jamais será vencido!

Agradeço a minhas amigas Mirella Guida (Mizinha) e Cristina Albert (Shun), que, não importa o que aconteça, estão ao meu lado (há anos), sendo grandes exemplos de pessoas incríveis e fortes.

Agradeço a Rafael Nunes, meu amor e amigo. Obrigada por todo o companheirismo, paciência e carinho. Me acompanhou durante todo o processo de desenvolvimento deste trabalho, me ajudando a ficar calma sempre que o código não funcionava corretamente e me ensinando a escrever claramente. Sem ele, eu nunca lembraria de colocar esse documento no formato justificado.

Agradeço, por fim, à minha família. Ao meu pai, Ricardo, que está sempre com um sorriso para me receber, não importa o meu nível de mal humor. À minha mãe, Ivana, me dando suporte em todos os momentos. À minha irmã, Marina, pela amizade e lições de vida. Ao meu irmão, Rico, por todas as conversas viajadas - e pelas sugestões de músicas.

Resumo

Rede social é um conceito que existe desde o século XX, definindo a estrutura da sociedade em forma de uma rede (Scott, 2017). Com a grande quantidade de dados sendo produzidos no mundo, principalmente através das redes sociais virtuais, como o Facebook, os relacionamentos entre usuários e diversas entidades podem ser mapeados para grafos seguindo o conceito de redes sociais. Sabendo disso e com o advento de algoritmos que extraem informações de diversos tipos sobre essas redes, é possível modelar os dados e aplicar tais algoritmos de forma que gere conhecimento. As técnicas de predição de *links* visam encontrar arestas que não existem em certo momento em um dado grafo, porém têm uma tendência de existirem futuramente. Este trabalho tem como objetivo abordar algumas dessas técnicas para analisar sua eficiência dada uma rede formada por informações de visitas de pessoas a lugares físicos. Para atingir o objetivo, foi necessário analisar e processar os dados de visita para a construção de uma rede que representasse bem as relações entre pessoas e lugares. A partir das redes construídas, foram aplicados experimentos utilizando técnicas de predição de *links* visando identificar possíveis visitas de usuários a lugares. Por fim, foi feita uma comparação dos resultados das diferentes abordagens e suas performances.

Palavras-chave: Predição de *links*, localização, visita, rede heterogênea.

Abstract

Social network is a concept that has existed since the twentieth century, defining the structure of society in the form of a network (Scott, 2017). Having a vast amount of data being produced worldwide, especially through virtual social networks such as Facebook, the relationships between users and various entities can be mapped to graphs following the concept of social networks. Knowing this and with the advent of algorithms that extract information of diverse types of networks, it is possible to model the data and to apply such algorithms in a way that generates knowledge. The link prediction techniques aim to find edges that do not exist at a given moment in a given graph, but have a tendency to exist in the future. This work aims to address some of these techniques to analyze their efficiency given a network formed by information from visits of people to physical places. In order to reach the objective, it was necessary to analyze and process the visit data for the construction of a network that represented well the relations between people and places. From the built networks, experiments were applied using link prediction techniques to identify possible visits from users to places. Finally, a comparison was made of the results of the different approaches and their performances.

Keywords: Prediction of links, location, visit, heterogeneous network.

Sumário

Capítulo 1

Introdução

O comportamento de usuários inseridos no contexto atual da internet, a partir de aplicativos, sites de relacionamento, etc, é uma rica fonte de dados que podem ser usados para gerar conhecimento sobre hábitos e interesses dos mesmos. Campanhas publicitárias, por exemplo, têm a necessidade de atingir consumidores que terão interesse no produto ou serviço oferecido, sendo um grande mercado que pode usufruir das informações providas por inteligências que as extraem hábitos e interesses de usuários para que melhorem a performance de suas ações.

O Youtube[1], por exemplo, utiliza o conteúdo do vídeo, bem como os interesse de busca dos usuários para fornecer anúncios interligados a eles[2], dessa forma a chance de o usuário realmente ter interesse naquele conteúdo é maior.

O objetivo deste trabalho é explorar dados geolocalizados com o intuito de construir uma rede e analisar a capacidade de predição de locais que as pessoas têm interesse em ir utilizando técnicas de redes sociais. Essas técnicas são utilizadas em vertentes de estudo que tem o objetivo de entender comportamentos sociais (Liben-Nowell, 2007).

Com a utilização de uma base de dados bruta constituída por comportamentos *off-line* de um conjunto de pessoas, foi realizado um estudo sobre ela, analisando o significado das informações presentes, bem como formas de utilizá-las nas aplicações de técnicas de predição de *links*, filtrando e limpando informações de baixa utilidade.

O trabalho envolveu, em seguida, a construção de uma rede heterogênea utilizando a base curada relacionando usuários e locais (entidades principais da rede), constituídos por visitas de pessoas a bares em Recife. A partir dela, foram analisadas as técnicas existentes e formas de aplicá-las, com o objetivo de verificar os resultados das predições.

Este documento está estruturado em forma de capítulos com o conteúdo distribuído da forma explicada a seguir. No capítulo 2, serão explorados os conceitos básicos sobre os assuntos abordados na pesquisa, como os conceitos de redes sociais e recomendação de ponto

[1] https://www.youtube.com/
[2] https://adwords.googleblog.com/2017/09/know-their-intention-get-their.html

de interesse, bem como o tema principal, predição de *links*, mostrando também algumas técnicas e forma de avaliação. No capítulo 3 será apresentado o conjunto de dados utilizados e o processo de desenvolvimento sobre ele, como filtragens realizadas e a construção da rede. O capítulo 4 apresenta a metodologia dos experimentos, os algoritmos aplicados e os resultados adquiridos a partir deles. Por fim, o capítulo 5 expõe a conclusão do trabalho e possíveis trabalhos futuros.

Capítulo 2

Fundamentos

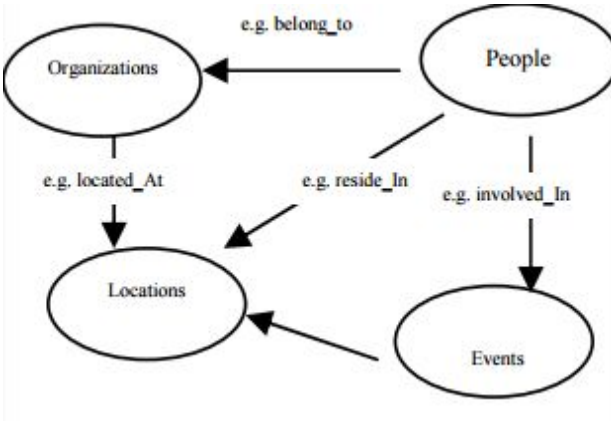
Neste capítulo serão apresentados importantes fundamentos para o entendimento do trabalho como um todo. Inicialmente, será explicado o conceito de rede social, suas aplicações e segmentações. Em seguida, a definição de recomendação de ponto de interesse e como a predição de *links* e suas técnicas podem atingir esse objetivo.

2.1 Redes sociais

Desde os primórdios da sociologia, estudiosos visualizavam a sociedade como organismos sociais, com estruturas e relacionamentos. Porém, apenas no século XX, os sociólogos alemães Simmel, Vierkandt e von Wiese criaram o conceito da estrutura social usando a analogia de uma rede ou *web* (teia), o que permitiu enfatizar a flexibilidade das redes sociais, em questão de estrutura e crescimento a partir de ações dos próprios membros da rede (Scott, 2017).

Redes sociais são constituídas de entidades e relacionamentos. Uma entidade é uma representação de um membro da rede, o qual pode possuir algum tipo de interação com outros membros, constituindo, assim, um relacionamento. As entidades e relações podem possuir diversas naturezas. Entidades podem representar pessoas, lugares, organizações, marcas, entre outros. Já relacionamentos representam predicados, por exemplo: “é amigo de”, “tem interesse em”, “visitou”.

Figura 1. Exemplo de entidades e relacionamentos



Fonte: Oellinger, 2006

A Figura 1 representa a estrutura básica de uma rede social, em que as entidades são Pessoas, Organizações, Lugares e Eventos. Nessa rede social, as pessoas *fazem parte de* organizações, *moram em* lugares e *estão envolvidas em* eventos. Ela pode ser um objeto de estudo para entender que tipos de eventos são frequentados por pessoas que fazem parte de certos tipos de organizações, por exemplo. Uma mesma rede social pode trazer diversos tipos de informações a depender da análise realizada.

É possível perceber que existem relacionamentos com diferentes granularidades que podem ser observados na rotina de uma pessoa. Essa granularidade pode ser enriquecedora para a análise realizada ou ter pouca utilidade. Dessa forma, entende-se que ao estruturar uma rede social para ser estudada, é necessário entender o tipo de análise a ser realizada para que a rede tenha apenas informações que serão úteis, especialmente porque técnicas de análise algorítmicas podem ter alta complexidade de tempo e memória sobre o tamanho da rede.

2.1.1 Rede social virtual

O termo "rede social" se tornou mais popularizado a partir da existência dos *sites* de relacionamento interpessoais, como Facebook[3], Twitter[4] e Linkedin[5]. Esses *sites* permitem que os membros participem pró-ativamente da construção da rede, em que os relacionamentos e interações são explicitamente declarados. Uma relação de amizade no Facebook consiste em um convite enviado de um participante para outro, o qual pode aceitá-lo ou recusá-lo. Já no Twitter, uma relação de interesse ocorre quando um participante

[3] https://www.facebook.com
[4] https://twitter.com/
[5] https://www.linkedin.com/

começa a *seguir* outro, levando-o a ver as informações expostas pelo segundo. Quanto mais tipos de interações e relações existirem nessa rede, mais complexa e completa ela será.

Essas redes podem ser exploradas para prover entendimento sobre os próprios participantes e seus comportamentos, podendo utilizar esse conhecimento para engajar os usuários em novas interações. No caso do Facebook, ele pode sugerir a amizade de duas pessoas ao perceber que elas têm um ou mais amigos em comum, ou sugerir um evento para algum usuário baseado em outros eventos que ele foi ou no fato de amigos desse usuário terem explicitado que vão para aquele evento. Esses exemplos mostram o tipo de ganho que pode ocorrer ao se estudar sobre redes sociais e técnicas de recomendação.

As redes sociais virtuais possuem altos números de usuários ativos, permitindo uma grande quantidade de dados sobre interações entre as entidades presentes nelas. O Facebook, por exemplo, possui mais de dois bilhões de usuários ativos por mês[6], sendo a maior plataforma de rede social virtual do mundo.

2.1.2 Rede social heterogênea

Dado o exemplo do Facebook, é possível identificar entidades de naturezas diferentes. Intuitivamente, a pessoa é a entidade principal da rede nesse caso, porém suas relações não se dão exclusivamente com outras pessoas. Existem relacionamentos entre pessoas e eventos, bandas, estilos de música, lugares, marcas, etc. Logo, o conjunto de entidades presentes nesta rede é diverso, bem como os tipos de relacionamentos. Nota-se, então, que uma rede não está limitada a apenas um tipo de entidade e relacionamento.

As redes que são constituídas por mais de uma entidade e/ou relacionamento de naturezas distintas são conhecidas como redes heterogêneas.

2.2 Recomendação de ponto de interesse

Pontos de interesse são locais, como lojas ou restaurantes, que possuem características atraentes para algum usuário. Essas características podem variar, indo desde o tipo de comida servida, por exemplo, até o fato de outro usuário frequentá-lo. A recomendação de pontos de interesse resolve o problema de usuários que desejam diminuir seu tempo de decisão ao escolher algum lugar para ir (Gao, 2015). Assim como ocorre na Netflix[7] com relação a

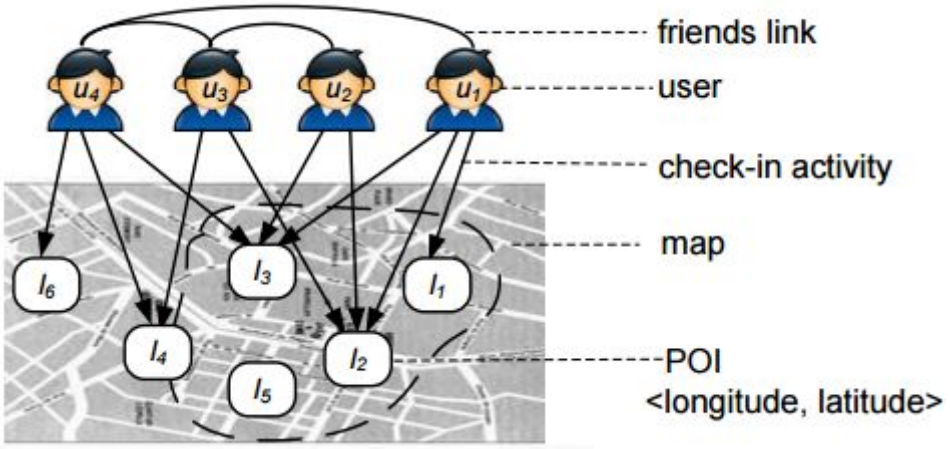
[6] https://newsroom.fb.com/company-info/
[7] https://www.netflix.com

filmes e séries, a recomendação de pontos de interesse pode ressaltar lugares que esse usuário provavelmente tem interesse em ir, dadas outras informações sobre ele e lugares existentes.

O leque de informações sobre lugares abrange quatro categorias principais: quem, quando, onde e o que. O "quem" se refere a mapear qual usuário foi àquele local. O "quando" se refere ao momento em que essa visita ocorreu, como um *timestamp*. O "onde" se refere ao local visitado. E o "o que" se refere a características intrínsecas ao local visitado, que podem contribuir para a recomendação como um todo (Gao, 2015).

Trabalhos recentes vêm explorando cada vez mais atributos nas redes sociais baseadas em localização, considerando os quatro Qs definidos acima. Essas redes são complexas pois, por serem heterogêneas, contém diversos tipos de arestas e nós, que talvez devam ser tratados de formas distintas na aplicação de algoritmos de recomendação.

Figura 2. Exemplo de rede com informações geográficas



Fonte: Ye, 2011

A Figura 2 mostra um exemplo de uma rede heterogênea. Os usuários possuem a relação de amizade conectando-os. Isso pode ser extraído de sites como o Facebook, por exemplo. As relações dos usuários com os locais se dão a partir de *check-ins*, os quais são registros de visita proativo, em que o usuário usa alguma rede como o Foursquare[8] para expor a informação de que está em certo local.

2.3 Predição de *links* temporais

A predição de *links* temporais em redes sociais parte do suposição de que a topologia da própria rede em um dado momento pode indicar quais serão os futuros *links* dela (Al

[8] https://www.foursquare.com

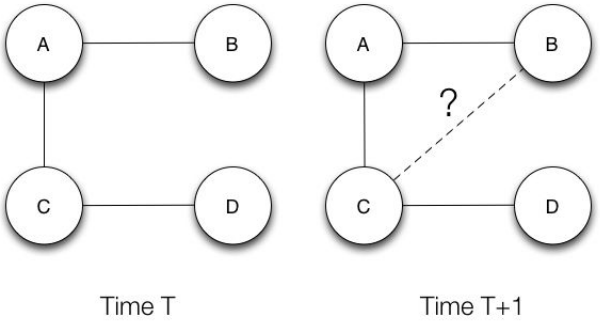
Hasan, 2011). Isso é possível ao transformarmos a rede em um grafo e aplicar técnicas sobre ele, como será explicado a seguir.

2.3.1 Definição do problema

Dado um grafo bidirecional, sem pesos nas arestas, $G =< V, E >$, cada aresta $e = (u, v) \in E$ representa um relacionamento entre os nós u e v que se deu em um momento *t(e)*. Dado os *timestamps* $t_0 < t_1 \leq t_2 < t_3$, o sub-grafo $G[t_0, t_1]$ contém todos os relacionamentos que ocorreram naquele intervalo temporal. Assim, é possível definir o resultado da predição de links como a lista de arestas que não estão presentes no sub-grafo $G[t_0, t_1]$, mas são prováveis de existir no sub-grafo $G[t_2, t_3]$.

A lista de arestas está ordenada de forma decrescente de acordo com um valor de similaridade (ou tendência) que pode ser calculado utilizando diversas técnicas conhecidas, como vizinhos comuns (*common neighbors*), que serão explanadas em seguida.

Figura 3. Exemplo do funcionamento de predição de *links* temporais



Fonte: https://engineering.linkedin.com/social-network-analysis/

A imagem 3 indica um exemplo da predição de links temporais, em que no tempo T existem apenas três links e foi previsto, de acordo com técnicas aplicadas, que em um momento T+1 há a possibilidade de existir uma aresta entre os nós C e B.

2.3.2 Técnicas

As técnicas de predição de *links* se baseiam bastante em teoria dos grafos e análise de redes sociais, podendo ser modificadas para o propósito em questão (Liben-Nowell, 2007). A convenção é que os métodos devem calcular, dado dois nós presentes na rede, um valor de similaridade ou proximidade entre eles. Esse valor pode ser interpretado como um grau de tendência de que dois nós vão se conectar no futuro, sendo utilizado para ranquear e comparar todas as arestas que o preditor fornece como sugestões.

Como os grafos podem representar redes sociais, as suposições feitas pelas técnicas são fortemente relacionadas a comportamentos reais na sociedade, com a ideia de que se duas pessoas têm vários amigos em comum, por exemplo, a chance delas se tornarem amigas é grande.

Técnicas baseadas em vizinhança utilizam a quantidade de vizinhos (nós que estão conectados por uma aresta) para avaliar a similaridade. O cálculo de *Common Neighbors* (vizinhos em comum) é o mais direto nessa abordagem, em que a similaridade entre dois nós da rede é dada pela quantidade de vizinhos que tais nós têm em comum.

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

Na fórmula acima, Γ denota o conjunto de vizinhos de um nó. Trabalhos como o de Newman (2001), obtiveram resultados positivos na utilização dessa técnica em uma rede construída a partir de colaborações em trabalhos científicos, mostrando a utilidade desse cálculo.

O coeficiente de Jaccard, bastante utilizado na área de recuperação de informação (Salton, 1983), mede a probabilidade de um elemento aleatório pertencente a união dos conjuntos de vizinhos dos dois nós estar na interseção dos mesmos, como mostra a fórmula abaixo.

$$similaridade(x,\ y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Adamic e Adar (2003) propuseram um cálculo de similaridade também baseado em características, porém ressaltando características mais raras. No contexto de vizinhança, isso é traduzido para: Dados os nós x e y, e um vizinho em comum z, quanto menos vizinhos z tem, mais significativas são suas conexões.

$$similaridadeAdamicAdar(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|}$$

Ao invés de utilizar a vizinhança diretamente, existem técnicas que fazem uso de caminhos no grafo para avaliar a similaridade (ou tendência de conexão) entre dois nós. A menor distância entre dois nós, por exemplo, é a maneira mais direta de utilização de caminhos, de forma que a similaridade é calculada da seguinte forma:

$$similaridadeMenorCaminho(x,y) = -|menorCaminho(x,y)|$$

Para que os menores caminhos sejam mais relevantes, a fórmula acima usa o valor negativo do tamanho do menor caminho entre os nós. Essa técnica é conhecida como *Graph Distance* (distância no grafo).

Apesar da simplicidade, o cálculo de menor caminho desconsidera características da topologia do grafo que podem ser relevantes para avaliar se dois nós realmente possuem uma tendência a se conectarem no futuro. Um exemplo é a quantidade de caminhos. Se existem cinco caminhos de tamanho dois entre os nós *x* e *y*, e apenas um caminho de tamanho dois entre *x* e *z*, de acordo com a fórmula de similaridade de menor caminho ambos os pares terão o mesmo valor, -2, porém, se for considerado que a quantidade de caminhos também é relevante, esse cálculo não seria robusto para o exemplo descrito e não iria diferenciar os dois pares.

O algoritmo de Katz (1953) utiliza todos os caminhos possíveis entre dois nós para calcular a tendência que eles têm de se conectarem, ressaltando os caminhos de cada tamanho individualmente através de um multiplicador. Isso significa que caminhos menores podem ter mais influência sobre o resultado final do que caminhos mais longos.

$$katz(x,y) = \sum_{l=1}^{\infty} \beta^l \bullet \left| caminhos^l(x,y) \right|$$

Na fórmula acima, β é o parâmetro de influência que vai diferenciar os pesos para cada tamanho de caminho possível. No caso dos caminhos, existem duas variantes possíveis. Primeiramente, é possível considerar valores binários, ou seja, se existe algum trajeto entre *x* e *y* com tamanho *l*, então o valor resultante da função *caminhos* será 1, caso contrário, será 0. Isso faz com que a quantidade de caminhos por tamanho não influencie no resultado. Por isso, a função *caminhos* também pode ser a quantidade de trajetórias de tamanho *l* entre *x* e *y*, sendo assim a segunda variante.

A *friends-measure* (medida de amigos) considera que quanto mais *links* conectam dois nós através de algum vizinho em comum, maior a chance deles se conectarem (Fire, 2011).

$$friends-measure(x,y) = \sum_{u \in \Gamma(x)} \sum_{v \in \Gamma(y)} \delta(u,v)$$

$$\delta(u,v) = 1, \; se \; u = v \; ou \; (u,v) \in E \; ou \; (v,u) \in E$$

Com as fórmulas acima, é possível analisar que, dado um grafo com arestas bidirecionais, a friends-measure pode ser um caso particular do algoritmo de Katz com valores binários, em que $\beta = 1$ e $l_{max} = 2$.

Por fim, existem técnicas que utilizam *random walk*[9] (passeio aleatório), como Hitting Time e Rooted PageRank. A primeira delas, para um par de nós *x* e *y*, se refere a quantidade de passos necessários para um passeio aleatório começando em *x* até alcançar *y*. Uma desvantagem desse algoritmo é que, mesmo que *x* e *y* sejam próximos, existe uma dependência em partes mais distantes do grafo.

A técnica de Rooted PageRank resolve o problema descrito acima a partir da aplicação de um *reset* periódico que permite que o passeio aleatório volte para o nó inicial com uma probabilidade fixa a cada passo. Dessa forma, partes mais distantes do grafo são raramente visitadas (Liben-Nowell, 2007).

Neste trabalho, foi utilizada uma modificação das técnicas de caminhos Katz e Friends-measure, explicada na seção 3.4.

2.3.3 Avaliação

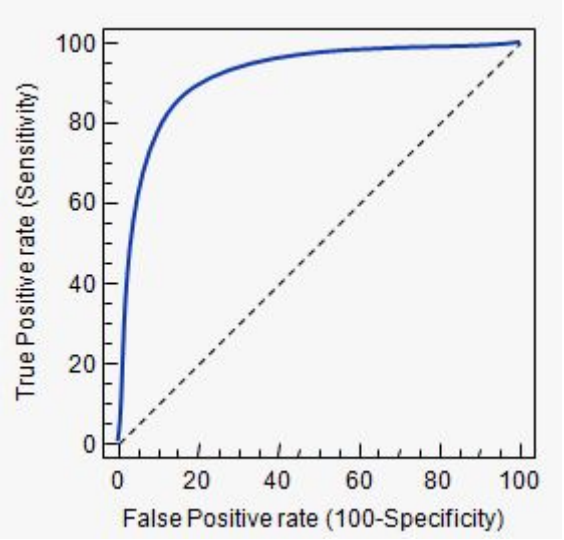
Para ser possível avaliar a qualidade de uma predição, existem metodologias de cálculo de métricas que permitem verificar os resultados das técnicas aplicadas. Como não é possível prever o futuro realmente, os dados utilizados precisam ser divididos em um conjunto de treinamento e de teste. As técnicas são aplicadas sobre o conjunto de treinamento, e então, o resultado será comparado de alguma forma ao conjunto de teste.

A geração dos conjuntos de treinamento e teste pode ser feita com a retirada de arestas relevantes da rede, de forma que, dado o grafo G = <V, E>, E' = E - R é o conjunto de arestas sem as arestas relevantes, que estão contidas em R. G' = <V, E'> é o conjunto de treinamento e o conjunto de teste é G.

A AUROC (*area under the receiver operating characteristics curve*) é uma métrica consolidada que representa a área abaixo da curva ROC. A curva ROC mostra o *trade-off* entre os benefícios (positivos verdadeiros) e os custos (falsos positivos), que pode ser visualizado em um gráfico bidimensional (Fawcett, 2005).

[9] https://en.wikipedia.org/wiki/Random_walk#Random_walk_on_graphs

Figura 4. Exemplo de curva ROC



Fonte: https://www.medcalc.org/manual/roc-curves.php

A Figura 4 mostra o gráfico que representa curva ROC, em que a linha pontilhada indica um preditor randômico e a linha azul indica um preditor mais assertivo. Quanto mais próximo ao canto esquerdo superior, melhor é o preditor utilizado.

No contexto de um ranking de valores ordenados, a AUROC é a probabilidade de um *link* relevante (que está presente no conjunto de teste e não está presente no conjunto de treinamento) ser ranqueado acima de um *link* irrelevante (que não está presente em nenhum dos conjuntos) (Wang, 2014). Dadas *n* comparações independentes, se existem *n'* ocorrências de *links* relevantes com um valor de similaridade maior que *links* irrelevantes e *n''* ocorrências de *links* relevantes e irrelevantes com o mesmo valor, a acurácia será equivalente a

$$AUC = \frac{n' + 0.5n''}{n}$$

Dadas distribuições normais, um AUC maior que 0.5 é melhor do que um preditor randômico (Wang, 2014). Quanto maior for o valor de AUC, melhor é o resultado da técnica aplicada.

Essa métrica permite um bom entendimento sobre a qualidade das técnicas aplicadas e será usada na análise de resultados do capítulo 4.

2.4 Trabalhos relacionados

Existem diversos trabalhos que envolvem aplicações de técnicas de predição de *links* em redes com heterogêneas e/ou envolvendo dados de localização para recomendação de ponto de interesse.

O trabalho realizado por Davis (2011) abordou a problemática de aplicar técnicas de predição de *links* em redes heterogêneas. Em um dos datasets estudados, a rede era bipartida, de forma que foi proposta uma alteração no algoritmo de Common Neighbors para redes heterogêneas, em que, dado dois nós x e y, é calculada a quantidade de caminhos de tamanho três entre x e y. Além disso, foi realizada uma comparação entre métodos supervisionados e não-supervisionados, mostrando resultados melhores dos métodos supervisionados.

O estudo realizado por Gao (2015) foi focado em adicionar informações de conteúdo sobre lugares na recomendação de ponto de interesse, com a intenção de enriquecer as informações de forma a melhorar a recomendação. Nesse trabalho, entretanto, não foram utilizadas técnicas de predição de *links*. Ao invés disso, foi utilizada fatorização de matrizes. Os resultados mostraram uma melhoria na qualidade de predições a partir do enriquecimento das informações.

No artigo de Scellato (2011), foi explorado o problema de formular um sistema de predição de *links* para redes baseadas em dados de localização, utilizando dados do Gowalla[10] , uma rede social em que os usuários podiam compartilhar suas localizações a partir de *check-ins*. O trabalho obteve resultados positivos, prevendo cerca de 66% dos *links* de interesse. A justificativa do bom resultado mostrou duas escolhas feitas pelos pesquisadores. A primeira delas foi focar a predição de *links* em um conjunto reduzido de pares de usuários, dado que a rede possui informação de relação de amizade entre eles. A segunda estratégia foi utilizar a atividade de *check-ins* do usuário para definir a relevância de um local para ele.

[10] https://en.wikipedia.org/wiki/Gowalla

Capítulo 3

Desenvolvimento

Este capítulo explora o conjunto de dados utilizados, mostrando as informações contidas nele e o processo de curadoria do mesmo para a criação da rede. Além disso, é abordada a estrutura das redes construídas e qual técnica foi selecionada, assim como as devidas modificações para que se aplicasse aos grafos.

3.1 Conjunto de dados

O conjunto de dados foi fornecido pela In Loco[11], startup recifense que trabalha com localização em diversas áreas, mantendo a privacidade dos usuários, de forma que não é possível traduzir o identificador interno para informações como o nome de uma pessoa. O conjunto contém informações que foram adquiridas a partir de dados de geolocalização coletados passivamente, ou seja, nenhum usuário explicitou qualquer informação a partir de um *site* ou aplicativo. Em muitos trabalhos nessa área, o conjunto de dados é extraído de alguma rede social virtual, como no caso de Gao (2015) que utilizou o Foursquare como fonte para os dados de localização, em que o usuário realiza um *check-in* quando deseja expor, por algum motivo, sua visita àquele local.

Logo, nota-se que neste trabalho os dados descrevem o comportamento *offline* do usuário, como ele interage no mundo físico sem precisar se engajar em uma ação *online*.

A primeira coleta de dados foi feita adquirindo visitas de cinco finais de semana (sexta-feira a domingo) nos meses de agosto e setembro, com o intuito de entender as informações presentes e testar formas de processamento. Foram selecionados finais de semana partindo do pressuposto que as pessoas tendem a sair para mais locais diferentes do usual. Geralmente, durante a semana as pessoas seguem uma rotina de trabalho e atividades fixas, enquanto que nos finais de semana existe mais tempo livre para visitar outros tipos de lugares.

[11] https://www.inlocomedia.com/

Cada visita possui uma assertividade associada ao local onde ela ocorreu, variando na escala “baixa”, “média” e “alta”, seguida de uma lista de possíveis locais. A assertividade possui um papel muito importante na utilidade daquela visita, pois indica a certeza de que aquela visita ocorreu em certo local a partir das informações existente sobre aquela visita.

Tabela 1. Informação de assertividade nas visitas

Usuário	Assertividade	Locais
X	Alta	A
Y	Média	A, B, C
Z	Baixa	A, B, C, D, E, F

Fonte: A autora

A Tabela 1 exemplifica os casos possíveis. No caso da primeira visita, o usuário X visitou o lugar A com alta assertividade, ou seja, existe uma certeza associada que é bastante alta. Em toda visita que possui alta assertividade, existe apenas um local associado. No caso da segunda visita, Y realizou uma visita, porém existe uma grau de dúvida se ela se deu A, B ou C. A assertividade média indica que existem poucos lugares possíveis, mas a quantidade de informação não foi suficiente para garantir em qual desses lugares a visita ocorreu. No último caso, Z realizou uma visita, entretanto existem várias possibilidades de onde ela pode ter ocorrido, logo a assertividade foi baixa.

Cada lugar presente na base de dados contém zero ou mais rótulos relativos a sua natureza, como “restaurante” ou “museu”, além de diversas informações interessantes, como *rating* e faixa de preço. Com a análise da amostra, foi possível notar que apenas as informações de rótulos eram presentes na maioria dos lugares, porém cerca de 40% da base possui *rating* e 15% possui informação de faixa de preço. Como essas duas informações estão muito esparsas, não foram consideradas no desenvolvimento. Para a realização deste trabalho, apenas os rótulos foram utilizados.

O segundo conjunto de dados utilizado é composto por informações de visitas de pessoas a lugares em Recife no período de 01 de setembro de 2017 a 31 de outubro de 2017,

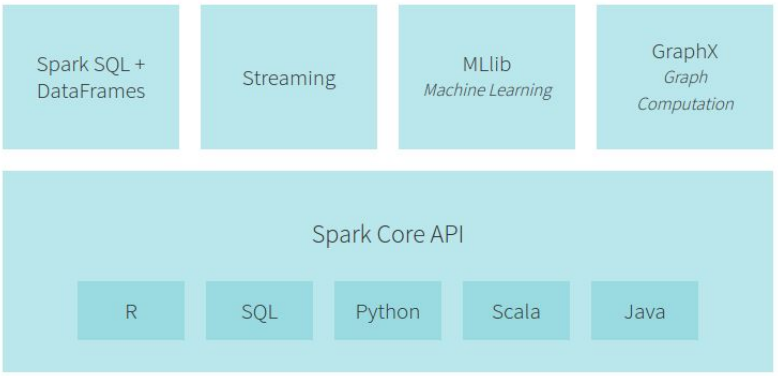
resultando em mais de 19 milhões visitas realizados por mais de 858 mil usuários únicos com uma base de mais de 60 mil lugares distintos.

Devido a grande quantidade de dados, o processamento dos mesmos se tornou bastante lento, necessitando de uma infraestrutura mais elaborada, o que levou a um custo adicional impeditivo na continuidade do trabalho. Além disso, a falta de especificidade dado o conjunto completo poderia acarretar em muito ruído, prejudicando resultados. Para diminuir a quantidade de dados com a menor perda de informação, foram aplicadas algumas heurísticas exploradas na seção seguinte.

3.2 Processamento dos dados

O Apache Spark[12] é um framework de processamento de Big Data que oferece velocidade e facilidade de uso. Possui módulos para processamento de dados em batch e streaming, podendo ler diversos tipos de dados, como parquet e csv, de maneira intuitiva. Apesar de ter um impacto maior ao rodar de forma distribuída, com alta performance, também pode rodar *standalone*, oferecendo também o Spark-shell, que provê uma interface de linha de comando interativa.

Figura 5. Módulos do Apache Spark



Fonte: https://databricks.com/spark/about

O Spark SQL é o módulo que permite trabalhar com dados estruturados, em que são utilizados DataFrames. Um DataFrame é uma coleção de dados organizada de forma colunar, em que cada coluna é tipada e possui um nome, se assemelhando a uma tabela de um banco de dados relacional. Essa organização dos dados permite que sejam processadas *queries* SQL, tanto no formato bruto, em que a query completa é passada como parâmetro, como através de métodos do Spark SQL.

[12] https://spark.apache.org/

Figura 6. Exemplo de utilização de Spark SQL

```
1  df.createOrReplaceTempView("users")
2  val users1 = spark.sql("select user_id from users")
3  val users2 = df.select("user_id")
```

Fonte: A autora

A Figura 6 mostra duas formas de aplicar o comando SELECT em um DataFrame, nesse caso, *df*, resultando no mesmo conteúdo nos DataFrames *users1* e *users2*.

Essas funções permitem realizar filtragens e extrair *insights* dos dados de forma simples, possibilitando a rápida avaliação de cada heurística aplicada no processamento do conjunto. Portanto, foi utilizada na fase de processar o conjunto de dados.

Figura 7. *Pipeline* de pré-processamento do conjunto de dados

Fonte: A autora

A Figura 7 mostra os passos feitos para processar a base de dados antes da aplicação dos algoritmos de predição de *links*.

Primeiramente, no passo 1, para evitar que informações nulas sejam tratadas como regulares, qualquer lugar ou usuário com identificador nulo foi removido da base, incluindo, consequentemente, a remoção das visitas que contém alguma informação nula, como a assertividade. Isso evita que dois lugares ou usuários sejam tratados com um único, quando na realidade são diferentes, o que adicionaria ruídos durante a aplicação dos algoritmos e poderia deteriorar a qualidade dos resultados.

Como as visitas não possuem nenhuma limitação quanto a qual classe de local foi visitado, os diversos rótulos presentes na base que caracterizam um local foram utilizados na filtragem. A princípio, existem 96 rótulos distintos, os quais representam categorias de diversos espectros. No espectro de transportes, por exemplo, existem os rótulos “estação de ônibus”, “ponto de táxi”, “aeroporto”, “posto de abastecimento”, “aluguel de carros”. Já no âmbito de compras, existem os rótulos “loja de livros”, “loja de roupas”, entre diversas outras.

Percebe-se, então, que existe uma falta de especificidade nos tipos de locais presentes na base de dados. Com o intuito de focar a análise em um tipo de local, foi selecionado o rótulo “bar”. Essa decisão foi baseada em alguns fatores importantes.

Primeiramente, a existência de várias visitas a lugares com esse rótulo - já o rótulo “drive thru”, por exemplo, possui apenas 38 visitas, podendo não ser suficiente para atingir resultados interessantes. Além disso, bares são locais de alta interação social, em que as pessoas, muitas vezes, vão para interagir com outras. Isso indica que o fator social e a presença de certas pessoas podem levar outras a visitarem esses locais. Esse tipo de interação é muito importante na aplicação de técnicas de redes sociais.

Para filtrar a base de visitas com apenas as que ocorreram em locais que possuem certo rótulo, foi necessário aplicar uma função *flat* para desacoplar visitas com mais de um local possível, como observado no exemplo abaixo.

Tabela 2. Visita com lista de lugares acoplada

Usuário	Assertividade	Locais
A	Média	X,Y,Z

Tabela 3. Visita com locais desacoplados

Usuário	Assertividade	Local
A	Média	X
A	Média	Y
A	Média	Z

Fonte: A autora

A Tabela 3 acima mostra o resultado da aplicação da função sobre a visita presente na Tabela 2, o que permite filtrar as visitas em cada local separadamente a partir dos rótulos selecionados. Se os locais X e Y possuem, entre seus rótulos, o rótulo “bar”, eles serão mantidos, enquanto a possível visita para o local Z, não tendo o rótulo "bar", seria descartada. Esse processo representa os passos 2 e 3 do *pipeline* de pré-processamento, na Figura 7.

O último filtro aplicado nesta etapa de preparação do dados foi remover todas as visitas com assertividade baixa. O grande volume de lugares possíveis em uma visita de assertividade baixa pode introduzir bastante ruído. Por exemplo, se uma visita de assertividade baixa possui 10 lugares possíveis, significa que seriam considerados nos algoritmos 9 relações que não ocorreram de fato. O impacto dessa categoria de visitas é potencializado dada a percentagem de visitas desse tipo com relação ao total, equivalente a 86.6%, resultando em mais da metade da base sendo formada por visitas que não ocorreram.

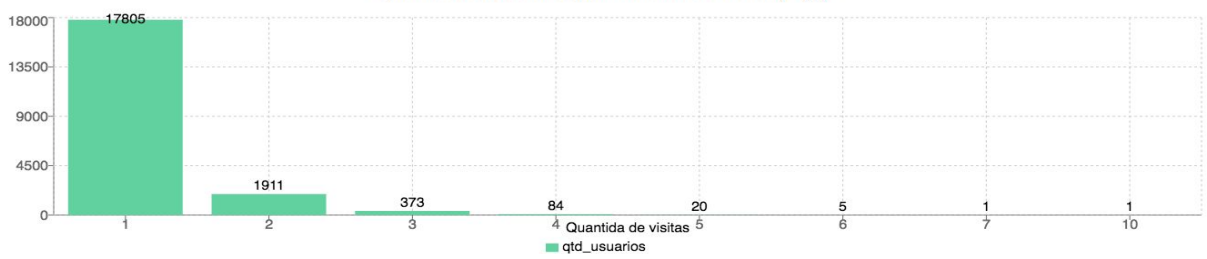
A aplicação desses filtros resultou em 23.229 visitas de assertividade alta e 58.345 de assertividade média. Dessa forma, foram formados dois conjuntos de dados: o primeiro possui apenas visitas de assertividade alta, com 20.200 usuários únicos e 226 lugares, já o segundo considera tanto as visitas de assertividade alta quanto de assertividade média, contendo 55.947 usuários únicos e 1.077 lugares.

3.3 Estrutura da rede

O processo de construção da rede formalizada na descrição do problema de predição de *links* (seção 2.3.1) ocorreu para cada um dos conjuntos de dados. Cada usuário é representado por um nó na rede, utilizando seu identificador, assim como cada lugar. Para cada visita V(u, l), foi adicionada uma aresta bidirecional entre *u* e *l*.

No primeiro conjunto de dados, considerando apenas visitas com alta assertividade, as arestas não possuem peso, ou seja, todas as visitas vão possuir uma influência idêntica na aplicação dos algoritmos, constituindo, assim, o Grafo 1. Como a quantidade de usuários únicos é muito próxima da quantidade de arestas, existem muitos usuário com apenas uma visita.

Figura 8. Gráfico da quantidade de usuários por quantidade de visitas (assertividade alta)



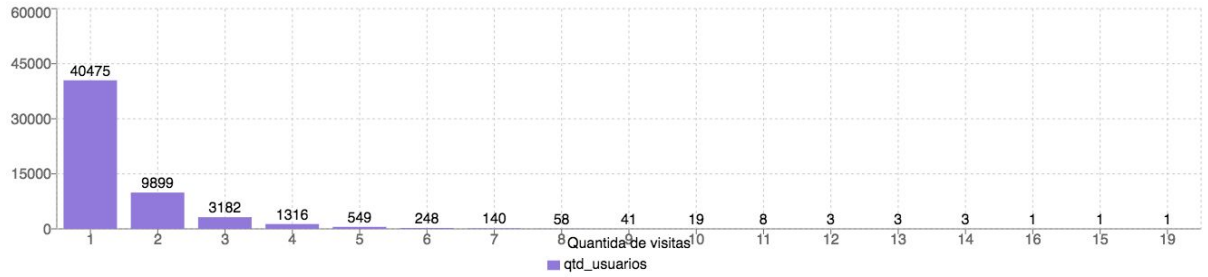
Fonte: A autora

O gráfico da Figura 8 mostra a quantidade de usuários distribuído pela quantidade de visitas que aquele usuário realizou. 88% dos usuários possuem apenas uma visita na base, o que é um indicativo comportamental dos usuários com relação a frequentarem bares. Essa informação pode ser traduzida para: Cerca de 17 mil pessoas frequentaram, com certeza, um bar nos meses de setembro e outubro de 2017.

No segundo conjunto de dados, foram construídas duas redes. A primeira delas foi através do mesmo processo descrito anteriormente, em que todas as arestas possuem o mesmo peso. Dessa forma, visitas de assertividade média e alta são tratadas da mesma forma, constituindo o Grafo 2. A segunda construção adiciona pesos nas arestas, de forma que se a

assertividade for alta, o peso será 1, e se for média o peso será 0.5. A seleção desses pesos foi arbitrária, dado que a inserção de incerteza nesse tipo de análise não possui um *benchmark*.

Figura 9. Gráfico da quantidade de usuários por quantidade de visitas (assertividade alta e média)

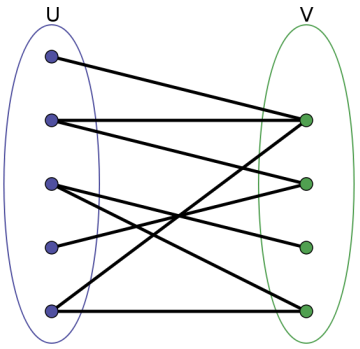


Fonte: A autora

Com o gráfico acima, observa-se que, apesar de existirem mais usuários únicos, existem mais usuários com mais visitas.

Os três grafos gerados possuem algumas características que são fundamentais para entender a aplicação dos algoritmos. Primeiramente, são bipartidos, o que significa que os nós do grafo podem ser divididos em dois conjuntos disjuntos, de forma que toda aresta interliga um nó de um conjunto a um nó do outro conjunto.

Figura 10. Exemplo de uma rede bipartida



Fonte: https://en.wikipedia.org/wiki/Bipartite_graph

A imagem 10 mostra um exemplo de um grafo bipartido. Neste trabalho, como um usuário pode visitar apenas lugares e lugares podem apenas ser visitados por usuários, um dos conjuntos é formado pelos nós que representam usuários (representado por U, na imagem) e o outro conjunto contém todos os lugares presentes no grafo (representado por V, na imagem).

A partir da explicação acima, percebe-se que cada conjunto é formado por tipos de nós diferentes, o que indica que a rede formada é heterogênea. Além dos nós serem diferentes, se observarmos as arestas bidirecionais, cada sentido dela possui um predicado diferente. Dado que o conjunto U representa o conjunto de usuários e V representa o conjunto

de lugares, o usuário $u \in U$ e o lugar $v \in V$, a aresta $e(u, v)$ tem a semântica "visitou", ou seja, o usuário *u visitou* o lugar *v*, enquanto a aresta $e(v, u)$ possui o predicado "foi visitado por", ou seja, o lugar *v foi visitado* pelo usuário *u*.

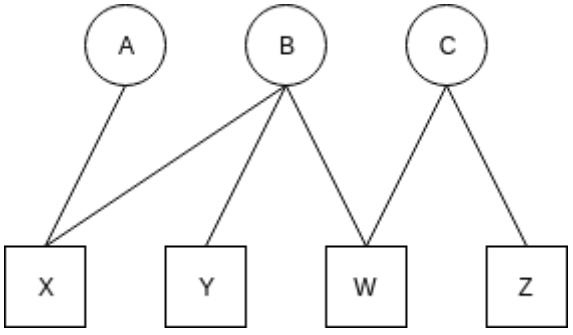
3.4 Algoritmo aplicado

A intenção da aplicação de técnicas de redes sociais é encontrar arestas que podem existir entre um nó que representa um usuário e um nó que representa um lugar.

As técnicas descritas na seção 2.3.2 se aplicam a grafos homogêneos, porém dadas as redes construídas neste trabalho, a aplicação dessas técnicas precisam ser modificadas. Técnicas que utilizam a vizinhança imediata, como *Common Neighbors*, Jaccard e Adamic/Adar, não funcionam diretamente. Isso ocorre por que, como um local apenas terá conexões com usuários e usuários apenas terão conexões com locais, nunca existirá uma intersecção entre os conjuntos de vizinhos de um local e um usuário.

Já com relação às técnicas relacionadas a caminhos, a aplicação pode ser a mesma, porém ignorando nos resultados as arestas que interligam nós do mesmo conjunto. Portanto, o algoritmo aplicado foi uma modificação de Katz, se assemelhando a Friends-measure.

Figura 11. Exemplo de grafo com as mesmas propriedades dos grafos criados



Fonte: A autora

A Figura 11 contém um exemplo de um grafo com as mesmas características dos grafos construídos. Dado que a intenção é sugerir lugares que o nó A tem a possibilidade de ir considerando os caminhos para outros nós, temos um caminho de tamanho três para Y e W ($A \rightarrow X \rightarrow B \rightarrow Y$ e $A \rightarrow X \rightarrow B \rightarrow W$), enquanto temos um caminho de tamanho 5 para Z ($A \rightarrow X \rightarrow B \rightarrow W \rightarrow C \rightarrow Z$) . Nas redes construídas, apenas caminhos de tamanho ímpar, maiores que um, irão conectar usuários a lugares que não estão previamente conectados.

Para poder analisar a influência dos menores caminhos, foi decidido considerar apenas a quantidade de caminhos de tamanho três, ou seja, a quantidade de caminhos do

menor tamanho possível para alcançar outro nó que faz parte do conjunto oposto. Por tanto, a fórmula de similaridade aplicada foi:

$$Similaridade(u, v) = |C3(u, v)|$$

Na fórmula acima, C3 é o conjunto de caminhos de tamanho três entre os nós *u* e *v*. A similaridade entre um usuário e um local pode ser interpretada como a tendência do usuário *u* ir ao local *v*.

Observa-se que o resultado será o mesmo que aplicar Katz com $\beta = 1$ e $l_{max} = 3$, utilizando a variante que considera a quantidade de caminhos entre dois nós, como explicado na seção 2.3.2. Esse cálculo será utilizado na realização dos experimentos do capítulo seguinte.

Para o processamento do grafo e aplicação do algoritmo, houve um processo de estudo sobre qual a melhor tecnologia para usar. O Spark oferece os módulos GraphX[13] e GraphFrames[14], ambos com métodos para realizar processamento sobre grafos. Os dois módulos possuem representações de grafos geradas a partir do tipo de dados que utilizam e são equivalentes com relação ao tipo de funções que podem ser aplicadas, como calcular a quantidade de tríades no grafo. Apesar dos benefícios de processamento de dados em grande quantidade e funções de auxílio, a utilização de Spark na aplicação dos experimentos se tornou desvantajosa quando colocada em comparação a utilização de outra linguagem.

Como o conjunto de dados estava mais conciso e sua estrutura requer a existência de modificações nos métodos existentes, as vantagens de utilizar um desses dois módulos de Spark seriam minimizadas. Portanto, foi escolhida a linguagem C++ para a aplicação dos algoritmos. A linguagem é altamente eficiente e a familiaridade com a mesma na construção de algoritmos que atuam sobre os grafos adicionou velocidade e facilidade nessa fase do desenvolvimento.

[13] http://spark.apache.org/graphx/
[14] https://docs.databricks.com/spark/latest/graph-analysis/graphframes/index.html

Capítulo 4

Experimentos

Este capítulo é constituído por duas seções que explicam os experimentos realizados. Primeiramente, a metodologia, ou seja, quais foram os passos seguidos para a realização dos experimentos. Então, na seção de resultados, será explicitado o que foi alcançado com a realização dos mesmos.

4.1 Metodologia

Para ser possível analisar a qualidade do algoritmo de predição, é necessário que exista uma forma de certificar que a aresta sugerida de fato é relevante. Com isso, foi selecionado um conjunto aleatório de duzentas arestas (cerca de 1% da quantidade total de arestas do Grafo 1) de assertividade alta, que foram removidas dos três grafos construídos. Assim, a intenção é analisar se essas arestas foram sugeridas e com que valor de similaridade quando comparadas a outras arestas sugeridas que não são relevantes. Para fins da terminologia desse experimento, as arestas que foram removidas serão chamadas de relevantes, enquanto arestas que forem sugeridas mas não fazem parte do conjunto removido são chamadas de irrelevantes.

Algoritmo 1. Pseudocódigo do algoritmo de predição para grafos sem pesos nas arestas

```
1  entrada: arestasRelevantes, grafoFiltrado
2  saída: aurocs
3  aurocs = {}
4  for (usuário, lugar) in arestasRelevantes do
5      comparaçõesPositivas = 0
6      comparações = 0
7      nósAlcançáveis = encontrarNósAlcançáveis(grafoFiltrado, usuário)
8      if nósAlcançáveis.vazio() then
9          continue
10     for lugarAlcançado in nósAlcançáveis do
11         if lugarAlcançado != lugar:
12             if Similaridade(usuário, lugar) > Similaridade(usuário, lugarAlcançado) then
13                 comparaçõesPositivas = comparaçõesPositivas + 1
14             else if Similaridade(usuário, lugar) == Similaridade(usuário, lugarAlcançado) then
15                 comparaçõesPositivas = comparaçõesPositivas + 0.5
16             comparações = comparações + 1
17     aurocs.inserir(comparaçõesPositivas / comparações)
18 retorne aurocs
```

Fonte: A autora

O Algoritmo 1 itera por todas as arestas relevantes, como mostrado na linha 4, encontrando, para cada usuário presente em uma aresta relevante, os nós alcançáveis em

caminhos de tamanho três (linha 7). Se não for possível alcançar nenhum nó a partir daquele usuário, significa que aquele nó estava isolado na rede. Isso é possível, por exemplo, quando o usuário tinha apenas uma visita; uma vez que a visita foi removida (se tornando uma aresta de interesse), o nó não possuía mais arestas de saída para iniciar um caminho possível. Como o algoritmo se baseia unicamente em caminhos, não será possível sugerir essa aresta.

Para cada nó atingível por pelo menos um caminho de tamanho três (linha 10), são comparadas as quantidades de caminhos que existem do usuário inicial e os nós de interesse e alcançável naquela iteração utilizando a fórmula descrita na seção 3.4. O resultado da comparação é critério para decidir a influência desses caminhos no cálculo da AUROC referente ao nó inicial (linhas 12 a 15).

Por fim, é retornado o conjunto de acurácias calculadas para todas as arestas relevantes naquele grafo.

Esse algoritmo foi aplicado para os grafos 1 e 2, nos quais todas as arestas possuem o mesmo peso.

Entretanto, no Grafo 3, a aplicação desse algoritmo não adiciona informações relevantes, pois como, nesse caso, existem pesos nas arestas, caminhos diferentes devem ter influências diferentes no cálculo, de forma que a quantidade de caminhos, por si, não faz essa distinção.

Tabela 4. Caminhos possíveis entre os nós A e X

	Aresta 1	Aresta 2	Aresta 3
Caminho 1	1	1	1
Caminho 2	1	1	1
Caminho 3	0,5	0,5	0,5

Tabela 5. Caminhos possíveis entre os nós A e Y

	Aresta 1	Aresta 2	Aresta 3
Caminho 1	1	1	0,5
Caminho 2	0,5	1	1
Caminho 3	1	0,5	1

Fonte: A autora

As tabelas 4 e 5 contém dois exemplos fictícios de caminhos encontrados no Grafo 3. Em ambas, existem três caminhos possíveis do usuário A para os respectivos lugares. A Tabela 4 mostra que dois caminhos são formados por arestas de visitas que de fato ocorreram, enquanto que no terceiro caminho existe incerteza em todas as arestas. Já na Tabela 5, em todos os caminhos existe um *link* atrelado a incerteza, porém dois são visitas de alta assertividade.

Para decidir, dado o exemplo, qual dos lugares deveria possuir um grau de tendência maior para o usuário A, foi levado em consideração que a incerteza pode ser um ruído, ou seja, ela pode significar que aquela aresta não existe de fato. Logo, para punir os caminhos que possuem incerteza, a relevância de um caminho é calculada pela multiplicação dos pesos das arestas. Assim, por quanto mais incerteza um caminho for formado, menor será a sua influência.

Com isso, cada caminho de tamanho três possui um valor associado. Para atribuir um valor de similaridade, ou tendência, entre o usuário e o lugar, é calculada a soma dos pesos de todos os caminhos.

$$Similaridade(u, v) = \sum_{c \in C3(u,v)} c(1) \cdot c(2) \cdot c(3)$$

A fórmula acima indica o cálculo final, de forma que para cada caminho entre dois nós, serão multiplicados os pesos de cada aresta e então somado ao resultado final.

Esse cálculo não é robusto com relação ao tamanho do conjunto de caminhos, se for desejado garantir que um par usuário-local com caminhos incertos fique abaixo de um par com caminhos sem incerteza.

Para dois locais l’ e l’’, se existe um caminho de l’ para o usuário u’ em que todas as arestas são construídas a partir de visitas de alta assertividade e existem 100 caminhos de l’’ para u’ em que todas as arestas foram construídas a partir de visitas de assertividade média, o grau de tendência entre l’ e u’ será 1, porém será 6,25 entre l’’ e u’.

Apesar do fato de que caminhos sem incerteza não possuírem ruídos, se a quantidade de caminhos incertos (formado por arestas de assertividade média) é alta, ela não deve ser desconsiderada. Isso foi considerado dada a suposição de que, mesmo que a visita seja incerta, ou seja, o usuário não necessariamente visitou aquele local, ele, com certeza, visitou algum local próximo, o que pode ser considerado um fator de influência.

Dessa forma, a soma das multiplicações foi utilizada como cálculo de similaridade entre usuário e lugares no algoritmo abaixo.

Algoritmo 2. Pseudocódigo do algoritmo de predição para grafos com pesos nas arestas

```
entrada: arestasRelevantes, grafoFiltrado
saída: aurocs
aurocs = {}
for (usuário, lugar) in arestasRelevantes do
    comparaçõesPositivas = 0
    comparações = 0
    nósAlcançáveis = encontrarNósAlcançáveis(grafoFiltrado, usuário)
    if nósAlcançáveis.vazio() then
        continue
    for lugarAlcançado in nósAlcançáveis do
        if lugarAlcançado != lugar:
            if somaMultiplicaçõesCaminhos(usuário, lugar) > somaMultiplicaçõesCaminhos(usuário, lugarAlcançado) then
                comparaçõesPositivas = comparaçõesPositivas + 1
            else if somaMultiplicaçõesCaminhos(usuário, lugar) == somaMultiplicaçõesCaminhos(usuário, lugarAlcançado) then
                comparaçõesPositivas = comparaçõesPositivas + 0.5
            comparações = comparações + 1
    aurocs.inserir(comparaçõesPositivas / comparações)
retorne aurocs
```

Fonte: A autora

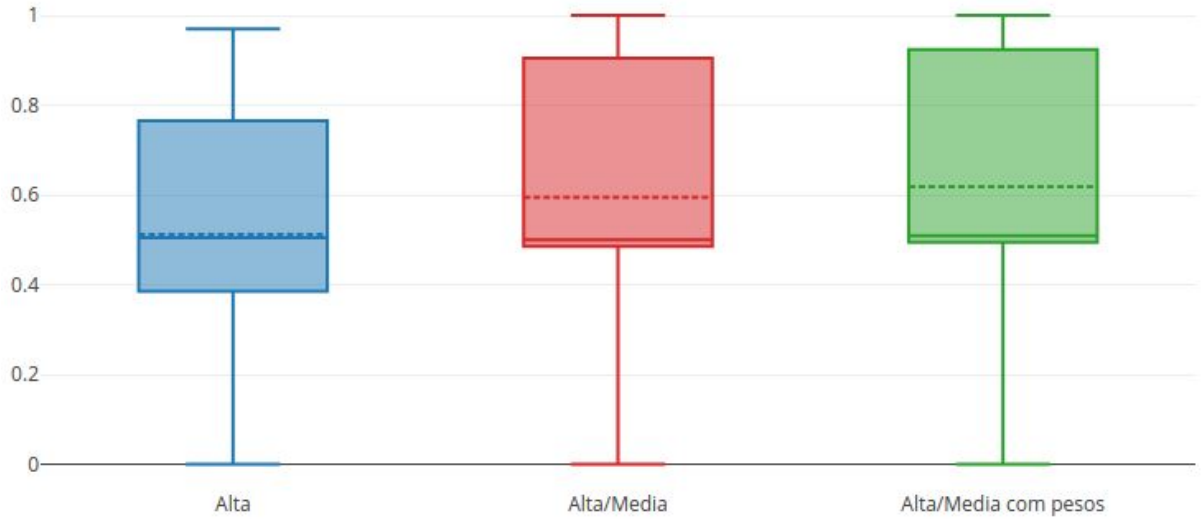
É possível observar que o Algoritmo 2 se diferencia do Algoritmo 1 apenas no processo de cálculo da similaridade. No algoritmo 2, os valores comparados são calculados da forma descrita anteriormente, em que é utilizada a soma das multiplicações dos pesos das arestas de cada caminho possível entre dois nós.

4.2 Resultados

Dado o conjunto de arestas relevantes, a quantidade de arestas que poderiam ser sugeridas foi de 26% no grafo com arestas de assertividade alta, e de 44% em ambos grafos com arestas de assertividade alta e média. Isso significa que, no melhor caso, 26% e 44% das arestas, respectivamente, poderiam ser preditas. Com a aplicação dos algoritmos, obteve-se uma cobertura de 23% no Grafo 1 e 38% nos Grafos 2 e 3, ou seja, para a maioria das arestas que poderiam ser preditas, existe pelo menos um caminho possível que conecta o usuário e local de cada aresta.

Utilizando os conjuntos de acurácias resultantes, foram geradas duas visualizações em formato de *boxplot* (diagrama de caixas) para comparação das distribuições.

Figura 12. Boxsplot dos valores de acurácia para cada aresta relevante



Fonte: A autora

O primeiro diagrama da Figura 12 (em azul) mostra os resultados do algoritmo aplicado sobre o grafo apenas formado por arestas de assertividade alta. A média dos valores, evidenciada pela linha pontilhada, foi 0.51. Esse resultado é muito próximo do valor de um preditor randômico, o que mostra uma qualidade ruim do preditor construído. Isso pode se dar ao fato da quantidade de visitas de alta assertividade não serem suficientes para que a topologia criada seja capaz de conter informações relevantes. Além disso, existe uma problemática do que realmente as arestas significam, que será discutido mais a frente nesta seção.

No caso do Grafo 2 (em vermelho), a inserção de arestas de assertividade média não deterioraram os resultados. Na verdade, essa melhora mostra que, apesar de existir ruído sendo adicionado ao grafo, ou seja, arestas que não realmente existem, as arestas adicionadas servem como um reforço para a consolidação de caminhos que levam a nós de interesse.

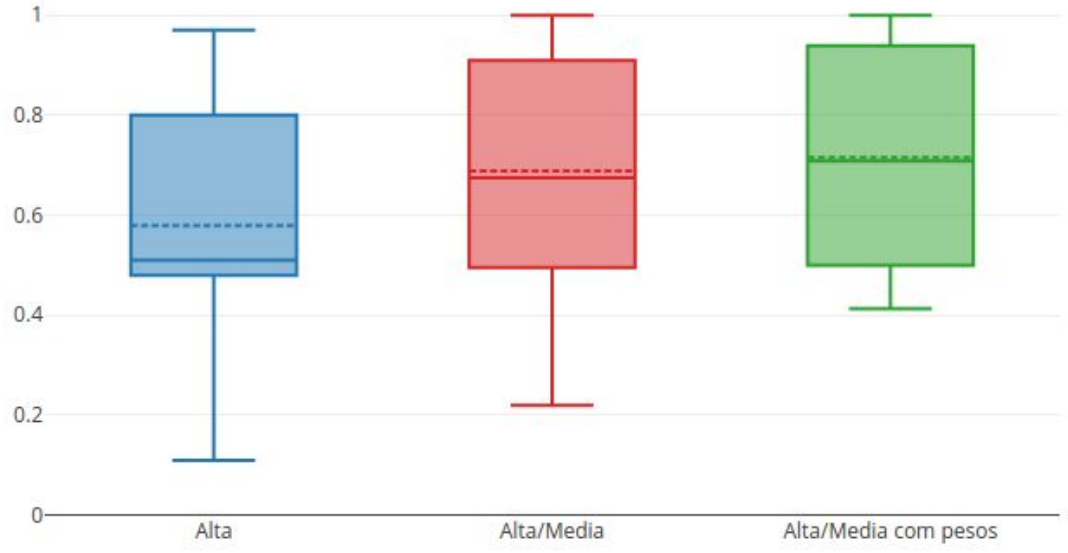
O último diagrama (em verde) mostra o resultado para o Grafo 3, formado por arestas de assertividade alta e média com pesos atribuídos 1 e 0.5, respectivamente. Nele, existe uma melhora quanto ao caso anterior, de forma que mostra que evidenciar os caminhos de alta assertividade, porém apoiados pelos caminhos que contêm incerteza, aumenta a confiança nos resultados, já que os ruídos não serão valorados com um grau de tendência alto.

A Figura 13, a seguir, considera apenas as arestas relevantes que foram sugeridas realmente, ou seja, foram ignorados os casos em que não existiu nenhum caminho possível entre os dois nós das arestas que foram removidas (arestas relevantes). A análise de alguns casos permite entender o motivo de não existirem caminhos.

Em alguns deles, dada uma aresta *(u,l)*, *l* (um lugar) possuía poucas visitas, o que diminuiu o espectro de usuários possíveis que pudessem visitar outros locais que *u* visitou. Nesse caso, existe uma dificuldade inerente à quantidade de dados para conseguir avaliar se a qualidade da predição foi o motivo para essa aresta não ter sido sugerida.

Além disso, é possível analisar que uma visita, apesar de ter assertividade alta, não possui informação sobre interesse, ou seja, apesar de a visita ter ocorrido, não há informações sobre se aquele usuário gostou ou não do local. Isso significa que o conjunto de arestas relevantes pode possuir ruídos, de forma que uma aresta considerada relevante para o algoritmo, não é relevante para o usuário. Com isso, o fato de algumas arestas não serem sugeridas pode indicar um bom sinal de que elas podem ser falsos-positivos no conjunto de arestas relevantes, não sendo de fato significativas para aquele usuário.

Figura 13. Boxsplot dos valores filtrados de acurácia para as arestas relevantes



Fonte: A autora

Na Figura 13, as médias das acurácias para os Grafos 1, 2 e 3 foram 0.58, 0.69 e 0.72, respectivamente. Esse resultado evidencia a existência de sugestões que estão entre as mais recomendadas para aquele nó, como pode ser visto observando os valores máximos e os terceiros quartis, os quais, em todos os casos, estão acima de 0.75. Os primeiros quartis

indicam que cerca de 75% das predições possuem uma acurácia acima de 0.5, sendo melhor que um preditor randômico nesses casos.

Esses resultados mostram que é possível fazer uso da topologia do grafo para encontrar arestas de interesse entre usuários e locais, entretanto várias situações decorrentes da origem dos dados e do que eles significam, além de lidar com incertezas, podem ser estudadas com o intuito da melhoria do preditor.

Capítulo 5

Conclusão

Neste trabalho, foram estudadas técnicas de predição de *links* sobre redes heterogêneas formadas a partir de dados baseados em localização. O conjunto de dados adquirido a partir da coleta de visitas de pessoas a lugares no mundo real, sem a interação ativa do usuário, permitiu captar mais atividades daquele usuário, ou seja, mais lugares frequentados por ele.

A inserção de incerteza, a partir de visitas nas quais existem várias possibilidades de locais onde ela ocorreu, fez com que existisse uma necessidade de adaptação e estudo sobre como utilizar essa incerteza para explorar todo o potencial da base.

Apesar de obter resultados positivos, existem vários possibilidades de trabalhos futuros, com questões interessantes a serem exploradas.

Por exemplo, a possibilidade de tratar as arestas como interesse e não como visitas. Nesse caso a relação entre um usuário e lugar seria construída a partir de um fator de recorrência e/ou levando consideração dados que possibilitem criar um perfil do usuário e relacioná-lo aos lugares frequentados. Além disso, outras formas de lidar com a incerteza das visitas pode traduzir melhor esse fator às técnicas de predição, além da avaliação.

O enriquecimento do grafo com mais informações sobre os usuários e os lugares, como dados demográficos, podem ser explorados para auxiliar na definição de interesse de um usuário por um local.

Além disso, aplicar as técnicas sobre grafos que possuem locais com outros espectros além de bares pode mostrar resultados diferentes, de acordo com o tipo de comportamento social que leva uma pessoa a visitar um local. Lojas de roupas, por exemplo; as pessoas podem visitar por recomendação de outras ou gostar das lojas que já conhecem e confiam. Essas questões podem ser analisadas utilizando a topologia da rede formada em algum trabalho futuro.

Outra questão interessante a ser explorada é a diferença da cultura em lugares distintos. Neste trabalho, foi analisado o comportamento em Recife, porém os resultados poderiam ser diferentes em um local como Nova Iorque, em que a vida noturna é muito ativa.

Dessa forma, observa-se que a utilização de uma base de dados com informações diversas e com a possibilidade de incerteza podem ser exploradas em diversos espectros, e este trabalho analisou um deles como forma de iniciar a pesquisa nesse conjunto.

Referências

Adamic, L. A., Adar, E. (2003). Friends and neighbors on the web. Social networks 25.3: 211-230.

Al Hasan, M., Zaki, M. (2011). Link prediction in social networks. Ebay Research labs, San Rose, California; Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York.

Davis, D., Lichtenwalter, R., Chawla, N. V. (2011). Multi-Relational Link Prediction in Heterogeneous Information Networks. Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana. Proceeding of the International Conference on Advances in Social Networks Analysis and Mining, Pages 281-288, Kaohsiung, Taiwan.

Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J., Cao, H. (2012). Link Prediction and Recommendation across Heterogeneous Social Networks. Department of Computer Science and Technology, Tsinghua University; Department of Computer Science and Engineering, University of Notre Dame; Nokia Research Center, Beijing.

Fawcett, T. (2005). An Introduction to ROC Analysis. Institute for the Study of Learning and Expertise, Palo Alto, California.

Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., & Elovici, Y. (2011, October). Link prediction in social networks using computationally efficient topological features. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 73-80). IEEE.

Gao, H.; Tang, J.; Hu, X.; Liu, H; (2015). Content-Aware Point of Interest Recommendation on Location-Based Social Networks. Computer Science and Engineering Arizona State University, Tempe, Arizona.

Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, 18(1), 39-43.

Liben-Nowell, D. and Kleinberg, J. (2007). The Link-Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology.

Newman, Mark EJ. "Clustering and preferential attachment in growing networks." Physical review E 64.2 (2001): 025102.

Oellinger, T., & Wennerberg, P. O. (2006). Ontology Based Modeling and Visualization of Social Networks for the Web. GI Jahrestagung (2), 94, 489-497.

Salton, G., McGill, M.J. (1983). Introduction to Modern Information Retrieval. New York.

Scellato, S., Noulas, A., Mascolo, C. (2011). Exploiting Place Features in Link Prediction on Location-based Social Networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1046-1054, San Diego, California, USA.

Scott, J. (2017, 4th edition). Social Network Analysis. SAGE publications.

Wang, F., Hu K., Tang Y. (2014). Robustness of Link-Prediction Algorithm Based on Similarity and Application to Biological Networks. Department of Physics and Key Laboratory of Intelligent Computing & Information Processing, Xiangtan University. Pages 246-252.

Ye, M., Yin, P., Lee, W., Lee, D. (2011). Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation. Department of Computer Science and Engineering, The Pennsylvania State University; Department of Computer Science and Engineering, Hong Kong University of Science and Technology.