Universidade Federal de Pernambuco – UFPE

Centro de Informática

Graduação em Ciência da Computação

Predição de links em uma rede heterogênea baseada em dados geolocalizados e de relacionamentos

Thiago Mota Bastos

Recife

2018

Thiago Mota Bastos

Predição de links em uma rede heterogênea baseada em dados geolocalizados e de relacionamentos

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.

Universidade Federal de Pernambuco – UFPE

Centro de Informática

Graduação em Ciência da Computação

Orientador: Prof. Ricardo Bastos Cavalcante Prudêncio

Recife

2018

Dedico este trabalho aos meus pais e a minha irmã por todo o suporte que me deram durante a graduação.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Paulo e Suylan, por sempre apoiar minhas decisões, dar conselhos nos momentos de dúvidas e por me formarem a pessoa que sou. Agradeço também a minha irmã, Marina, pela amizade e pelos momentos de gordices!

Também agradeço ao meu orientador, Ricardo Prudêncio, pela orientação no trabalho, pelas discussões de ideias e problemas durante o desenvolvimento.

Agradeço a In Loco por permitir utilizar o conjunto de dados nos experimentos desse trabalho e pela oportunidade de trabalhar em um lugar incrível que me faz crescer tanto profissionalmente como pessoalmente.

Agradeço aos amigos que fiz na faculdade, Danilo, Lapprand, Calegario, Lasalvia, Milena, Walber, João e Bormann, sem o companheirismo deles o caminho até aqui teria sido muito mais difícil.

Por fim, agradeço ao Estudando a Sociedade: Ozzy, Lari, Victor, Jenni, Marquinho, Pedrinho e Igor. Não importa o que aconteça, vocês estão sempre ao meu lado sendo esses amigos incríveis!

RESUMO

Rede social é um conceito antigo criado por sociólogos e antropologistas que estudavam a interligação social. A partir daí surgiu a ideia de redes sociais virtuais, como o Facebook, que deixa explícito o relacionamento entre usuários e outros atores, como empresas. Com os dados dessas redes é possível criar um grafo e aplicar algoritmos de predição de links para recomendar amizades ou até mesmo locais para visitar. Esse trabalho explora a predição de pontos de interesse usando dados geolocalizados e com informações de relacionamentos entre usuários. Para isso são utilizados dados de visitas e de conexões com redes wifi colhidas passivamente, ou seja, sem necessitar da ação direta do usuário informando o acontecimento. Com esses dados, foi criada um grafo com usuários, locais e redes wifis, além dos relacionamentos entre esses atores. A partir desse grafo, foi aplicado algumas técnicas de predição de links afim de analisar se há alguma melhora no desempenho ao utilizar os dados de relacionamento para recomendar pontos de interesse. Por fim, foi comparado o resultado de cada uma das abordagens e seus desempenhos.

Palavras-chave: rede heterogênea, predição de links, dados geolocalizados, dados de relacionamento.

ABSTRACT

Social networks is an old concept created by sociologists and anthropologists who studied the social interweaving. After this point, came the idea of virtual social networks, like Facebook, that makes explicit the relationship between users and other actors, like companies. With the data in this networks it's possible to create a graph and apply link prediction algorithms to recommend friendships or even places to visit. This work explores the prediction of points of interest using geolocalized data aswell as information about relationship of users. To achieve this, it's used data from visits and wifi conenctions gathered passively, that means without the necessity of direct action of the user notifying what is going on. With these data, it was created a graph with users, places and wifi networks, besides the relationship between these actors. With this graph, it was applied some techniques of link prediction to analyze if there is any performance improvement when using relationship data to recommend points of interest. Finally, a comparison was made of the results of each approach and its performance.

Keywords: heterogeneous network, link prediction, geolocalized data, relationship data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO

Atualmente as pessoas compartilham, principalmente nas redes sociais e aplicativos, todo tipo de dados sobre elas como interesses, locais onde esteve e círculo de amizades. Todos esses dados podem ser utilizados para gerar um perfil do usuário com seus interesses e hábitos, muito utilizado pelas campanhas publicitárias, por exemplo, para fazer campanhas voltadas a públicos muito específicos. Um exemplo pode ser pessoas que frequentam shoppings e gostam de animais. Além disso, é possível utilizar esse dado para trazer uma melhor experiência para o usuário, por exemplo em sites de livrarias, que utiliza as compras anteriores dos usuários para recomendar novos livros que eles possam se interessar.

Um exemplo da relevância desses sistemas de recomendações está no desafio proposto pela Netflix[1], que consistia em um prêmio de 1 milhão de dólares para quem conseguisse ter o melhor algoritmo de predição da nota que um usuário iria dar para um filme, baseado apenas em suas notas passadas.

1.2 OBJETIVO

O objetivo desse trabalho é explorar dados geolocalizados e de relacionamentos com o intuito de criar uma rede e analisar a capacidade de predição de locais de interesse da pessoa utilizando técnicas de redes sociais.

Foi realizado uma análise de uma base de dados constituída de visitas de usuários a locais e conexão dos usuários a redes wifis, buscando uma forma de inferir relacionamentos entre usuários com esses dados e como as técnicas de predição de link podem ser aplicadas a essa base, além de filtrar os dados para a viabilidade da execução.

Após isso foi construído uma rede heterogênea utilizando esses dados filtrados, relacionando usuários com locais e usuários com redes wifis. A partir disso foi definido algumas modificações das técnicas conhecidas para a aplicação nessa base.

Por fim foi comparado o desempenho de cada técnica, analisando se a utilização do conceito de relacionamento entre usuários obteve alguma melhoria em relação a já conhecida predição de locais baseados em locais passados.

[1] https://en.wikipedia.org/wiki/Netflix_Prize

1.3 ESTRUTURA DO TRABALHO

A estrutura do trabalho divide-se em capítulos. No capítulo 2 é abordado os conceitos fundamentais para o entendimento do trabalho, como o conceito de redes sociais, de pontos de interesse, de homofilia e o próprio conceito de predição de link, assim como apresentado as técnicas mais comuns. No capítulo 3 é apresentado o formato dos dados utilizados, o processamento realizado para tratar os dados para serem utilizados no trabalho, como a rede foi formada a partir desses dados e quais foram as técnicas utilizadas para os experimentos. O capítulo 4 é focado nos experimentos, primeiro explicando como são gerados os casos de teste e como eles serão avaliados, após isso é feito duas rodadas de experimentos, comparando o desempenho de cada técnica. Por fim, o capítulo 5 mostra as conclusões do trabalho e os possíveis trabalhos futuros.

2 FUNDAMENTOS

Nesse capítulo serão apresentados os fundamentos importantes para o entendimento do trabalho. Primeiro será explicado o conceito de redes sociais e redes sociais virtuais. Depois disso será discutido a predição de links, primeiro explicando o conceito de pontos de interesse, depois explicando sobre homofilia e, por fim, definindo o problema de predição de link e discutindo algumas técnicas comumente utilizadas. Além disso também há uma análise dos trabalhos relacionados.

2.1 REDES SOCIAIS

As redes sociais surgiram dos sociólogos e antropologistas que exploravam o entrelaçamento e a interligação das ações sociais utilizando metáforas como *tecido* ou *teia* da vida social. A partir disso, surgiram vários cientistas que começaram a usar a matemática para investigar a densidade e as conexões dessas redes, surgindo assim os principais conceitos de redes sociais (SCOTT, 2017).

Essas redes são constituídas de atores e relacionamentos. Os atores são as entidades sociais, ou seja, pode ser um indivíduo, uma empresa, um lugar, um grupo social, entre vários outros exemplos. Já os relacionamentos são as ligações entre os atores, essas ligações podem ser de vários tipos, como uma ligação que denota amizade entre dois indivíduos ou uma ligação de parceria comercial entre duas empresas (WASSERMAN; FAUST, 1994).

As redes não necessariamente são compostas por apenas um tipo de ator, por exemplo, pode haver uma rede formada por pessoas e locais, no qual um relacionamento entre uma pessoa e um local pode ter o significado de que aquele pessoa visitou o local. Não só isso mas também pode haver mais de um tipo de relacionamento, pegando o mesmo exemplo anterior, pode também haver uma conexão entre duas pessoas indicando que há uma amizade entre elas. Essas redes são chamadas de **Redes Heterogêneas**.

A Figura 2.1 representa uma estrutura básica de uma rede social, com os atores sendo as *Pessoas*, as *Organizações*, os *Lugares* e os *Eventos*, já os relacionamentos são descritos como: uma Pessoa *mora* em um Lugar, *participa* de um Evento e *faz parte* de uma Organização, já uma Organização *é localizada* em um Lugar.

Uma mesma rede social pode trazer diversar informações diferentes, dependendo da análise a ser feita. Por exemplo, com essa rede é possível extrair informações como o perfil das pessoas baseado no lugar que mora ou nos eventos que frequenta.

A quantidade de relacionamentos de um ator na rede também pode ser utilizado como uma fonte de estudo. Por exemplo, uma marca de carros pode pegar todas as pessoas

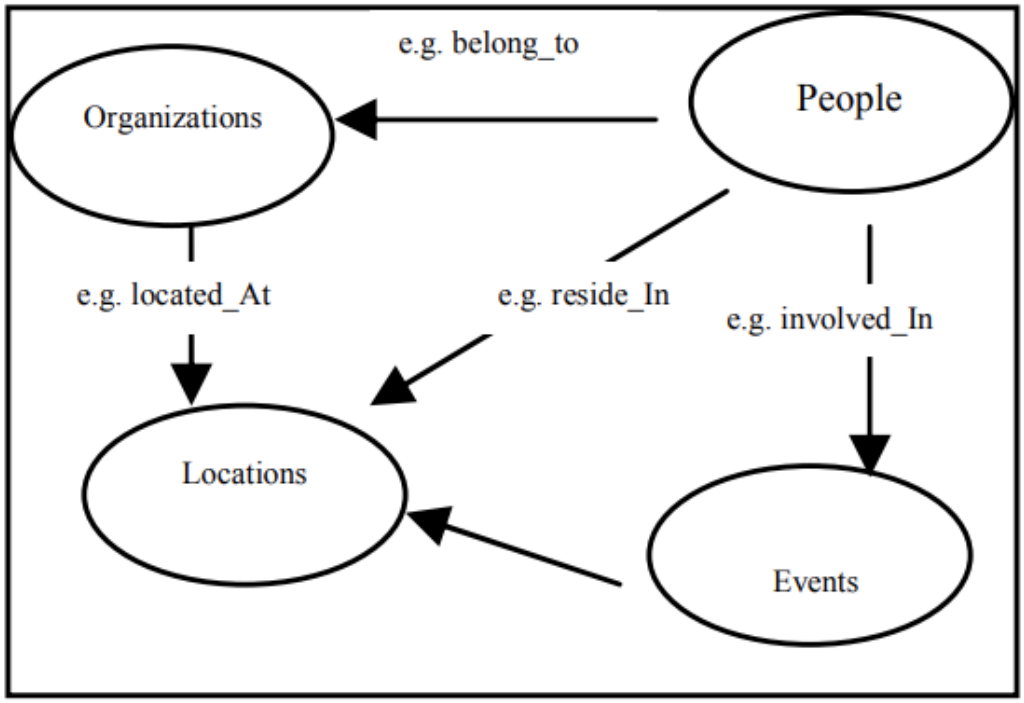

Figura 2.1 – Exemplo de uma rede social com vários tipos de atores e relacionamentos.
Fonte: (OELLINGER; WENNERBERG, 2006).

que falam sobre carros na internet e, a partir das conexões que cada pessoa tem, definir quem é mais influente e fazendo alguma publicidade com esse influenciador.

Portanto uma rede social pode ser uma rede muito complexa, com vários tipos de atores e relacionamentos que nem sempre são úteis para uma determinada análise, sendo assim é importante definir qual a análise que quer ser feita e quais são os dados necessários para ela, visto que as técnicas de análise podem ter um alto custo de desempenho dependendo do tamanho e das características da rede.

2.1.1 REDE SOCIAL VIRTUAL

O termo rede social ganhou popularidade com o advento dos sites de mídia social, como o Facebook[1] e o Twitter[2]. Esses sites facilitam a visualização da configuração de redes sociais pois seus usuários explicitam seus relacionamentos com outras pessoas, lugares ou organizações, por exemplo. Essa informação explícita, no Facebook, se dá pelo pedido de amizade que um usuário pode fazer a outro que, se aceito, configura uma relação de amizade entre os dois, já no Twitter essa ligação é mais direcional, onde um usuário pode seguir outro sem precisar ser seguido de volta, especialmente interessante para a descoberta de influenciadores na rede, ou seja, pessoas que possuem muitos seguidores.

[1] https://www.facebook.com/
[2] https://twitter.com/

Essa rede pode ser utilizada para entender cada usuário, extraindo seu perfil e interesses, e utilizando esse conhecimento para o engajamento do usuário. No caso do Facebook, ele pode sugerir novas amizades ou um evento que você pode se interessar. No caso do LinkedIn[3], uma rede voltada para o mercado de trabalho, pode até sugerir vagas que estão abertas que combinem com o seu perfil. Esses exemplos mostram o ganho que o estudo dessas redes e de técnicas de recomendação podem trazer para o usuário.

2.2 PREDIÇÃO DE LINKS

2.2.1 PONTOS DE INTERESSE

Pontos de interesse são locais, como restaurantes, lojas e hotéis, que são interessantes para alguma pessoa. Esse interesse pode ser dado por diversos fatores, como pelos interesses do usuário serem compatíveis com o que o local fornece. A recomendação de pontos de interesse visa diminuir o tempo de escolha de um local para um usuário ir ao filtrar locais que não sejam de interesse dele (GAO et al., 2015).

Os trabalhos de recomendação de pontos de interesse vêm explorando cada vez mais atributos nas redes sociais baseadas em localização, como a informação de amizade entre um usuário e outro ou as características dos pontos de interesse, tentando melhorar a recomendação. Por exemplo, se um usuário visita um ponto de interesse que é descrito como um restaurante de comida vegetariana, é possível inferir que aquele usuário tem interesse em outros locais que sirvam esse tipo de comida. Porém isso trás uma complexidade a redes que, por serem heterogêneas, possuem mais de um tipo de ator e relacionamentos que talvez tenham que ser tratados de forma diferente pelo algoritmo de recomendação.

A Figura 2.2 mostra um exemplo de uma rede heterogênea que possui a informação de relacionamento entre usuários, de visitas desses usuários a lugares e até mesmo informação de qual foi a ordem que os usuários fizeram as visitas. Apesar desses dados, juntos, possuírem dados relevantes que possam ser usados para recomendação, eles também podem ser tratados como subredes distintas, podendo até pegar os dados de cada rede de fontes diferentes, mostrando que é possível pegar dados diferentes de várias fontes e juntar tudo para fazer um recomendador mais complexo e, possivelmente, melhor.

2.2.2 HOMOFILIA

Outras informações para recomendação, como as pessoas que os usuários se relacionam, não teriam sentido impactar nos interesses do usuário se não fosse o princípio sociológico da homofilia.

[3] https://www.linkedin.com/

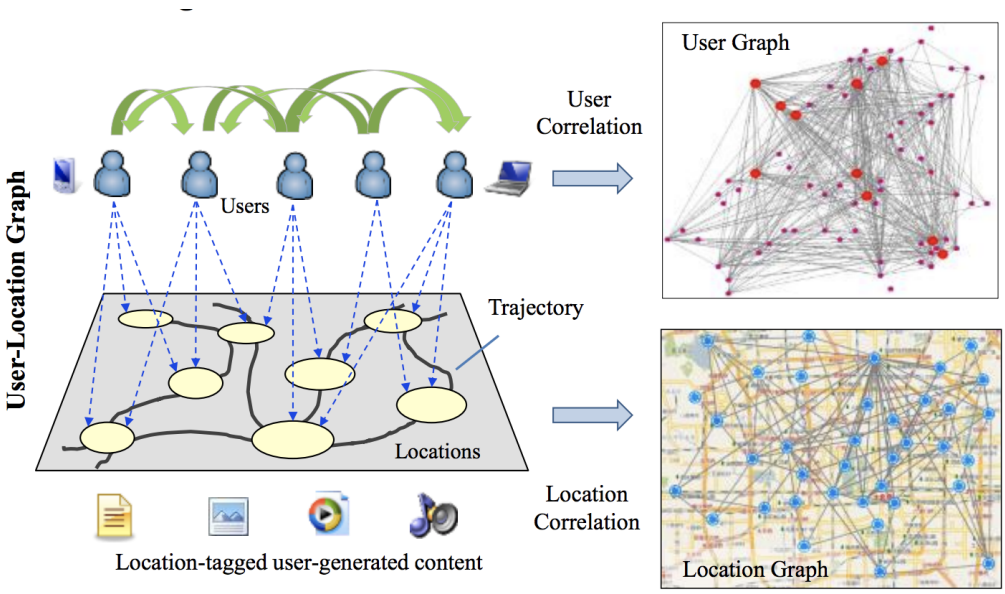

Figura 2.2 – Exemplo de uma rede que utiliza diferentes tipos de dados.
Fonte: (ZHENG, 2011).

Homofilia é o princípio que o contato entre pessoas parecidas ocorre mais frequentemente do que o contato de pessoas que não são parecidas. Isso implica que a distância em termos de características sociais se traduz em distância na rede, ou seja, em um número maior de relacionamentos que uma informação deve percorrer para conectar dois indivíduos (MCPHERSON; SMITH-LOVIN; COOK, 2001).

Sendo assim, como pessoas parecidas estão mais próximas na rede social, os algoritmos de recomendação conseguem aproveitar esse princípio para utilizar o dado de amizade entre usuários e fazer uma recomendação melhor, visto que lugares que os amigos de um usuário já visitou já possuem uma maior propensão que o usuário vá visitar aquele local.

2.2.3 DEFINIÇÃO DO PROBLEMA

A predição de links parte da suposição que a topologia da rede em um dado momento pode indicar quais serão os futuros links que serão formados entre os atores (HASAN; ZAKI, 2011). Isso é possível ao transformar a rede em um grafo e aplicar técnicas de recomendação nele. Podemos transformar a rede como um grafo através da seguinte definição:

Dado um conjunto de nós V, um conjunto de arestas E e um grafo bidirecional $G(V, E)$, no qual uma aresta $e = (u, v) \in E$ representa alguma interação entre os nós u e v. Dado os timestamps $t_0 < t_1 \leq t_2 < t_3$, o subgrafo $G[t_0, t_1]$ contém todos os relacionamentos que ocorreram nesse intervalo temporal. Assim a tarefa de predição pode ser definida como a lista de arestas que não estão presentes em $G[t_0, t_1]$, mas que podem aparecer no subgrafo $G[t_2, t_3]$.

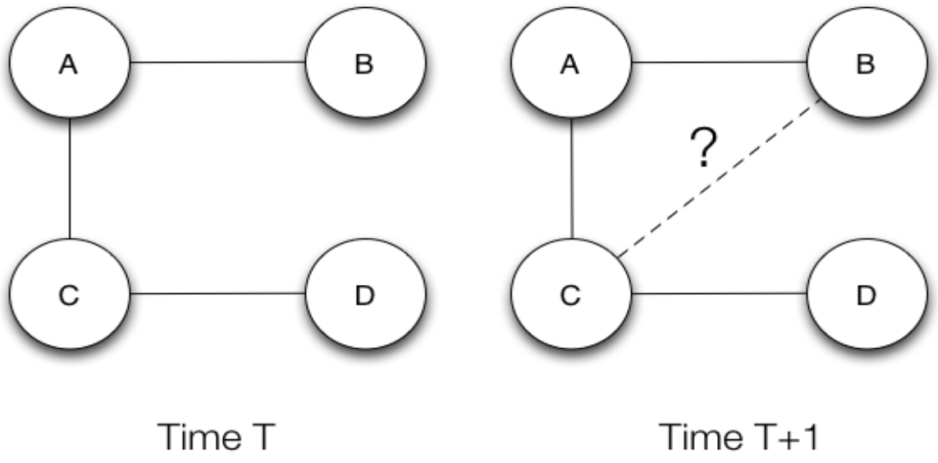

Figura 2.3 – Exemplo de predição de link.
Fonte: https://engineering.linkedin.com/social-network-analysis/organizational-overlap-social-networks-and-its-applications.

A Figura 2.3 mostra um exemplo de predição de link onde, no instante T, há apenas as arestas entre *A*-*B*, *A*-*C* e *C*-*D*, e, segundo as técnicas aplicadas de recomendação, foi previsto um link entre *C*-*B* no instante T+1.

2.2.4 TÉCNICAS

Em geral, todas os algoritmos calculam uma pontuação para um possível relacionamento entre dois nós, a partir do grafo, e gera uma lista ordenada da maior pontuação até a menor. Sendo assim os algoritmos podem ser vistos como uma computação de proximidade ou similaridade entre dois nós. No geral, todos os métodos são adaptações de técnicas utilizadas em teoria dos grafos e em análise de redes sociais (LIBEN-NOWELL; KLEINBERG, 2007).

No método de *Common Neighbors* (vizinhos em comum), a sua pontuação pode ser definida pela Equação 2.1. Onde $\Gamma(x)$ significa todos os vizinhos do nó x, ou seja, todos os nós que tem uma ligação com x.

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2.1}$$

A ideia por trás desse método é que, se um nó x tem uma conexão com um nó z e um nó y tem uma conexão com o mesmo nó z, tem uma probabilidade alta de que também há uma conexão entre x e y.

Apesar de simples, (NEWMAN, 2001) obteve resultados positivos na utilização dessa técnica em uma rede construída a partir de colaborações em trabalhos científicos.

Como a técnica de vizinhos em comum não é normalizada, há a técnica conhecida como *Coeficiente de Jaccard* que faz essa normalização como mostrado na Equação 2.2.

$$Jaccard(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2.2}$$

A ideia desse coeficiente é a probabilidade de que um nó que esteja na interseção dos conjuntos de vizinhos do nó x e do nó y sejam selecionados se for feita uma seleção aleatória de um nó na união do conjunto de vizinhos desses dois nós.

Já *Adamic e Adar* (ADAMIC; ADAR, 2003) propuseram uma técnica cuja pontuação é dada pela Equação 2.3.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|} \tag{2.3}$$

O objetivo desse cálculo é dar um peso maior ao vizinho em comum que possui um grau pequeno, ou seja, quanto menos conexões um vizinho em comum tem, mais significativa são suas conexões.

Por fim, o algoritmo de Katz (KATZ, 1953) utiliza todos os caminhos possíveis entre dois nós, multiplicando os caminhos de cada tamanho por um fator β. Isso significa que caminhos menores podem ter mais influência sobre o resultado final, como mostra a Equação 2.4.

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l . |caminhos^l(x, y)| \tag{2.4}$$

2.3 TRABALHOS RELACIONADOS

Há diversos trabalhos que tratam sobre recomendação em redes heterogêneas e sobre recomendação envolvendo dados de localização.

O trabalho de (DAVIS; LICHTENWALTER; CHAWLA, 2011) abordou a predição de links em uma rede heterogênea, mais especificamente uma rede bipartida de doenças e genes. Ele sugeriu uma modificação da técnica de vizinhos em comum que, em vez de só contar a quantidade de vizinhos em comum, ele deu um peso para cada vizinho baseado na contagem da ocorrência daquela tríade na rede. Além dessa técnica, foram utilizadas mais algumas não-supervisionadas e também algumas técnicas supervisionadas, que se saíram melhor no final.

Já (GAO et al., 2015) abordou o tema de recomendação de pontos de interesse. A abordagem desse trabalho foi adicionar informações de sentimento e de interesse dos usuários, retirado de comentários dos mesmos em redes sociais baseadas em localização como

o Foursquare[4], além dos dados do próprio ponto de interesse para fazer a recomendação. A técnica usada foi bem diferente, não utilizando nenhuma técnica de predição de link propriamente dita, mas sim utilizando de fatorização de matrizes. Os resultados encontrados foram melhores ao enriquecer os dados com essas informações de sentimento e interesse dos usuários do que sem.

No trabalho de (SCELLATO; NOULAS; MASCOLO, 2011), é abordado a predição de link baseado em dados de localização retirados de uma rede social baseada em localização, onde os usuários podem fazer *check-in* nos lugares que visitaram, essa rede foi a da Gowalla[5]. A técnica utilizada para predição foi a predição supervisionada utilizando as técnicas não-supervisionadas como características para o algoritmo supervisionado. No final obteve resultados positivos que, segundo os pesquisadores, foi consequência de duas escolhas feitas por eles. A primeira sendo em focar em um conjunto reduzido de pares de usuários, já que a rede possui informação de amizade entre eles. A outra escolha foi a de explorar os *check-ins* do usuário para definir a relevância de um local para ele.

[4] https://pt.foursquare.com/
[5] https://en.wikipedia.org/wiki/Gowalla

3 DESENVOLVIMENTO

Este capítulo explora o conjunto de dados, mostrando as informações contidas nele e os processamentos realizados para torná-lo apto a ser utilizado no trabalho. Além disso é mostrada a estrutura da rede construída e definido os algoritmos que serão utilizados nos experimentos, com suas devidas técnicas e modificações para lidar com esse tipo de rede.

3.1 CONJUNTO DE DADOS

O conjunto de dados foi fornecido pela startup recifense In Loco[1], que utiliza dados geolocalizados colhidos através de aplicativos parceiros com objetivo de gerar tráfego de pessoas para lojas anunciando em sua plataforma de anúncios ou gerar engajamento para seus aplicativos parceiros.

Os dados são coletados passivamente pelo celular quando um usuário visita algum local. Nesses dados coletados, além de possuir a informação do identificador único daquele local, também encontra-se o estado da conexão do usuário com redes WIFI, ou seja, se ele está conectado a alguma rede e as informações dela, em caso positivo.

Com esses dados dos lugares que os usuários visitaram, consegue-se extrair o comportamento do usuário no mundo físico, indicando quais são seus gostos a partir dos locais que frequenta. Já com o dado de conexão WIFI, extrai-se a informação de relacionamento, visto que, se poucas pessoas se conectaram a uma mesma rede, é um forte indicativo de que aquelas pessoas tem algum relacionamento relevante, como moram na mesma casa ou visitam a casa um do outro, ao contrário de pegar uma rede em que milhares de pessoas se conectam, por exemplo as fornecidas por restaurantes ou shoppings, em que provavelmente as pessoas que estão conectadas a elas nem se conhecem.

A partir desse dado de relacionamento, é possível utilizar o conceito de homofilia para fazer recomendações de lugares, nesse caso específico, mas não limitando-se a isso, obtendo resultados muito mais acertivos explorando um universo de possibilidades muito menor.

3.2 PROCESSAMENTO DOS DADOS

Para a realização desse experimento, foram coletados todos os dados de visitas entre os dias 1 de Setembro de 2018 até o dia 30 de Setembro de 2018. Porém todo esses dias correspondem a 3 terabytes de dados, fazendo o processamento desses dados ser inviável,

[1] https://inloco.com.br/

necessitando de uma infraestrutura mais elaborada, elevando o custo para a continuidade do trabalho. Sendo assim, para diminuir a quantidade de dados sem perder informação, foram aplicadas algumas heurísticas mostradas na Figura 3.1.

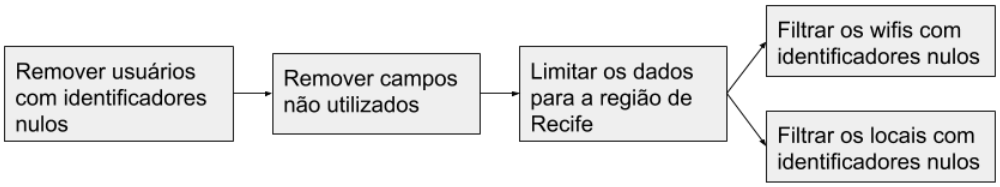

Figura 3.1 – Etapas do processamento realizado nos dados.

1. Remover usuários com identificadores nulos

   Na lógica de extração dos dados da In Loco, é utilizado um identificador único disponibilizado pelos dois sistemas operacionais para celulares, o iOS e o Android. Porém, o usuário pode não habilitar que esse dado esteja disponível para os aplicativos instalados, sendo assim o identificador, por padrão, chega no formato *00000000-0000-0000-0000-000000000000*. Isso é um problema pois mais de um usuário vai possuir o mesmo identificador, ou seja, um mesmo identificador terá o comportamento de várias pessoas com perfis potencialmente diferentes, provocando um ruído na recomendação.

   Outra possibilidade também é ocorrer um erro no processamento dos dados e o identificador simplesmente não vir, o que causa o mesmo problema citado acima.

2. Remover campos não utilizados

   Como esse dado é usado com vários propósitos diferentes, há vários campos desnecessários para o escopo desse trabalho e que ocupam muito espaço como dados do aplicativo e dados usados internamente pelas heurísticas de detecção de visita.

3. Limitar os dados para a região de Recife

   Com o objetivo de diminuir a quantidade de dados para tornar o processamento viável, os dados foram limitados para visitas ocorridas apenas na cidade do Recife, visto que já representa um subconjunto dos dados relevante para o trabalho, com cerca de 470 mil usuários que possuem dados de visita e de wifi, em volta de 4 milhões e 723 mil visitas de usuários a locais distintos, 50 mil locais distintos e um total de 2 milhões e 277 mil conexões de usuários a wifis distintos.

4. Filtrar os wifis com identificadores nulos

   Assim como nos identificadores dos usuários, os pontos de acesso de wifi também possuem um identificador único, chamado de *BSSID*, porém ele também pode vir nulo, o que resultaria em uma wifi correspondendo a vários outros, afetando a inferência de relacionamentos que será descrita posteriormente.

5. Filtrar os locais com identificadores nulos

   Os locais também possuem identificadores únicos e que também podem vir nulo, causando o mesmo problema de um local corresponder a vários outros, dessa vez afetando a quantidade de caminhos alcançáveis a partir de um local, gerando ruído na recomendação.

Como as informações de conexão wifi e de locais visitados vêm do mesmo dado, os dois filtros de identificadores têm o mesmo conjunto de entrada.

Para realizar essas filtragens, foi utilizado o Apache Spark[2], um framework de processamento de Big Data que oferece velocidade e facilidade de uso. Possui como foco rodar de forma distribuída, mas também pode rodar em apenas um computador. Sua vantagem é a de possuir a habilidade ler uma gama de diferentes tipos de arquivos, como csv, parquet e avro, de maneira fácil e intuitiva.

Outro ponto do Spark que facilita esse processamento é o de permitir trabalhar com dados estruturados, utilizando os *DataFrames*, que se assemelham a uma tabela de um banco de dados relacional, já que os dados são organizados de maneira colunar, com cada coluna tipada e com um nome específico. Com essa funcionalidade, há a possibilidade de implementar as filtragens de uma maneira fácil, se assemelhando a SQL.

3.3 ESTRUTURA DA REDE

Com os dados processados, é possível formar um grafo como o mostrado na Figura 3.2.

Esse grafo possui três tipos diferentes de nós, sendo eles os nós que representam as wifis, os nós que representam os usuários e os nós que representam os locais.

As arestas possuem um significado diferente dependendo quais nós são ligados, se uma aresta conecta um nó de usuário com um nó de wifi, isso significa que aquele usuário se conectou pelo menos uma vez naquela wifi. Já quando a aresta conecta um usuário com um local, significa que um usuário visitou aquele local pelo menos uma vez.

Vale ressaltar algumas características desse grafo formado, sendo elas:

1. Não há arestas entre um wifi e um local;
2. Não há arestas conectando nós do mesmo tipo, por exemplo, não há uma conexão usuário-usuário.

Com essas características, nota-se que é um grafo tripartido, o que exigirá certas adequações nos algoritmos de recomendação que serão utilizados.

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

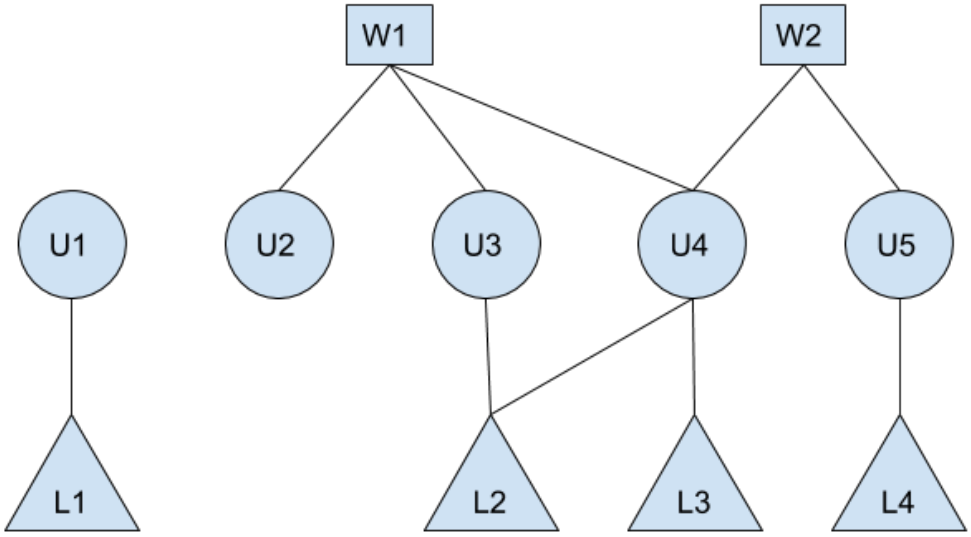

Figura 3.2 – Exemplo de grafo formado, onde os retângulos representam wifis, os círculos representam usuários e os triângulos representam locais.

Além disso, é possível extrair alguns dados analisando os dois subgrafos bipartidos que existem, sendo eles o grafo wifis-usuários e o grafo usuários-locais. Um dado relevante para ser analisado é o da esparsidade dos dois subgrafos, mostrados na Tabela 3.1.

Tipo de Nó	Número de Nós	Número de Conexões	Preenchimento do Grafo
Wifi	674.128	2.277.683	0,00048%
Locais	46.993	4.723.405	0,014%

Tabela 3.1 – Análise dos subgrafos wifis-usuários e locais-usuários

O cálculo do preenchimento do grafo foi feito utilizando a Equação 3.1:

$$Preenchimento = \frac{Número\,de\,Conexões}{Número\,de\,Nós \times Número\,de\,Usuários} \tag{3.1}$$

Como a quantidade de usuários no grafo é de 705.721, obtêm-se os dados de preenchimento demonstrados e percebe-se que o subgrafo de locais-usuários é um grafo bem menos esparso que o de wifis-usuários, o que poderá afetar negativamente algoritmos que usem os dados de wifi.

3.4 ALGORITMOS APLICADOS

Para fazer as recomendações foram escolhidos dois algoritmos bem simples para servir como base de comparação, o *Popular Place Recommender* e o *Popular Path Re-*

commender. Além desses dois algoritmos, foi feita uma modificação no Adamic-Adar para tornar possível aplicá-lo nesse grafo tripartido, o que resultou em dois algoritmos, o *Relationship Recommender* e o *Adamic-Adar Places Recommender*.

Todos os algoritmos de recomendações utilizados recebem uma lista de locais candidatos e devem retornar essa lista ordenada baseada em alguma heurística. A obtenção dessa lista inicial de locais candidatos será explicada no próximo capítulo, que irá falar dos experimentos.

3.4.1 POPULAR PLACE RECOMMENDER

Para o *Popular Place Recommender*, dado um local **l** qualquer, sua pontuação é definida pelo seu grau, como mostra a Equação 3.2.

$$Pontuação = deg(l),\ l \in Locais \tag{3.2}$$

3.4.2 POPULAR PATH RECOMMENDER

O *Popular Path Recommender* é uma variação do Katz, descrito na subseção 2.2.4, em que apenas são considerados caminhos de tamanho 3 e o β é igual a 1. Ou seja, dado um local **l** qualquer e um usuário **u**, a pontuação do local é definida pela quantidade de caminhos de tamanho 3 diferentes entre o usuário **u** e o local **l**, como mostra a Equação 3.3.

$$Pontuação = |C3(u,l)|,\ l \in Locais\, e\, u \in Usuários \tag{3.3}$$

Para os caminhos de tamanho 3, é considerado apenas o subgrafo de locais-usuários. Um exemplo de um caminho 3 por esse subgrafo pode ser visto na Figura 3.2, onde entre o usuário **U3** e o local **L3** tem um caminho definido por **U3 - L2 - U4 - L3**.

Esse algoritmo tem como finalidade pontuar um local para um usuário baseado nos locais que o usuário já visitou, considerando assim o perfil de visita do usuário, ao contrário do *Popular Path Recommender* que não faz essa consideração.

3.4.3 RELATIONSHIP RECOMMENDER

Para definir um relacionamento, foi considerado a conexão com wifis, onde, se duas pessoas se conectam a um mesmo wifi, elas têm algum tipo de relacionamento, por exemplo familiares que moram na mesma casa ou amigos que visitam a casa uns dos outros. Porém também é possível que seja uma wifi de um restaurante ou shopping, em que milhares de pessoas se conectam e isso não significa necessariamente que essas pessoas possuem algum tipo de contato mais íntimo.

Dado essa definição de relacionamento, percebe-se que se assemelha bastante ao uso do Adamic-Adar, com a exceção de se tratar de um grafo tripartido e esse tipo de recomendação é para nós do mesmo tipo.

Sendo assim, para o *Relationship Recommender*, dado um local **l** qualquer e um usuário **u**, a pontuação do local é definida usando o conceito de caminhos de tamanho 3, utilizado no *Popular Path Recommender* que foi modificado do Katz, e por uma modificação do Adamic-Adar, que normalmente é utilizado somando o inverso do logaritmo do grau dos vizinhos em comum de dois nós.

Como foi dito anteriormente, o Adamic-Adar irá focar em recomendar uma conexão entre nós do mesmo tipo, sendo assim utilizado para dar um peso ao relacionamento entre dois usuários, então, ao contrário do *Popular Path Recommender* que todos os caminhos de tamanho 3 tem o mesmo peso e usam o subgrafo locais-usuários, o algoritmo de relacionamentos utiliza do subgrafo wifis-usuários para dar um peso para o relacionamento entre usuários que irá repercurtir no peso dos caminhos para os locais que passam por aquele relacionamento. Ou seja, no final, apenas os caminhos que seguem a ordem **Usuário - Wifi - Usuário - Local** serão considerados, para conseguir aplicar o peso do relacionamento através do wifi.

Portanto sua pontuação é dada pela Equação 3.4.

$$Pontuação = \sum_{c \in C3(u,l)} relacionamento(u,v),\ l \in Locais, u\,e\,v \in Usuários\,e\,v \in c \quad (3.4)$$

Já a parte de relacionamento é calculado como mostrado na Equação 3.5. Onde o peso do relacionamento é dado pelo inverso do logaritmo do grau dos wifis que os usuários tem em comum.

$$relacionamento(u,v) = \sum_{c \in C2(u,v)} \frac{1}{log(deg(w))},\ w \in Wifis, u\,e\,v \in Usuários\,e\,w \in c \quad (3.5)$$

3.4.4 ADAMIC-ADAR PLACES RECOMMENDER

Parecido com o *Relationship Recommender*, o *Adamic-Adar Places Recommender* segue o mesmo princípio de aplicação do Adamic-Adar, porém, em vez de ser utilizado para cálculo de peso dos relacionamentos baseados em wifi, é utilizado para dar um peso a semelhança de perfis de usuários baseado em locais visitados. Ou seja, se poucas pessoas visitaram uma localização, é provável que as que visitaram possuam interesses em comum que valem a pena ser considerados na hora da recomendação.

Também diferentemente da recomendação baseada em relacionamento, e se aproximando mais do *Popular Path Recommender*, os caminhos que são considerados para esse

algoritmo são os que seguem a ordem **Usuário - Local - Usuário - Local**. Sendo assim temos a Equação 3.6 e a Equação 3.7.

$$Pontuação = \sum_{c \in C3(u,l)} semelhanca\_perfil(u,v),\ l \in Locais, u\,e\,v \in Usuários\,e\,v \in c \tag{3.6}$$

$$semelhanca\_perfil(u,v) = \sum_{c \in C2(u,v)} \frac{1}{log(deg(l))},\ l \in Locais, u\,e\,v \in Usuários\,e\,l \in c \tag{3.7}$$

4 EXPERIMENTOS

Este capítulo é constituído por duas seções que visam explicar os experimentos realizados. Na primeira seção é explicado a metodologia adotada, explicando como foram gerados os casos de teste e como os algoritmos foram comparados. Já na seção de resultados, é mostrado os resultados alcançados e as análises resultantes.

4.1 METODOLOGIA

Para a realização dos experimentos, foi utilizado o algoritmo proposto em (CREMONESI; KOREN; TURRIN, 2010), onde para cada caso de teste, é selecionado um usuário aleatório e, a partir desse usuário, é selecionado um dos locais que ele visitou e aleatorizados mais mil lugares que esse usuário não visitou. Sendo assim, no final temos um usuário com mil lugares não relevantes e um lugar relevante. Para esse trabalho foram gerados 1000 casos de testes seguindo essa regra.

Para conseguir comparar a qualidade de cada algoritmo de recomendação, é retirada a conexão do usuário com o local relevante selecionado e, depois do algoritmo fazer a recomendação e retornar a lista ordenada, é checado se o local relevante está entre os N primeiros, desse jeito calcula-se, para cada N, a sensibilidade *(recall)* e a precisão *(precision)*.

O recall pode ser definido como a quantidade de itens relevantes retornado pela recomendação sobre a quantidade de itens verdadeiramente relevantes. Sendo assim, como há **T** casos de teste, cada um com apenas um item verdadeiramente relevante, basta saber quantos itens relevantes foram retornados pela recomendação, que é justamente o **#acertos** na Equação 4.1. Ou seja, considerando os top **N** resultados de cada caso de teste, basta contar em quantos casos o local verdadeiramente relevante está contido nesse subconjunto, obtendo assim o cálculo do **recall**.

$$recall(N) = \frac{\#acertos}{|T|} \tag{4.1}$$

Já a precisão pode ser definida como a quantidade de itens relevantes retornado pela recomendação sobre a quantidade de itens recomendados. Seguindo a mesma lógica do **#acertos** descrito no cálculo do *recall*, percebe-se que o denominador é a quantidade de casos de teste **T** multiplicado pelo **N** sendo considerado, que dá a quantidade de locais que foram recomendados, resultando na Equação 4.2, que também pode ser escrita em

função do *recall*.

$$precision(N) = \frac{\#acertos}{N.|T|} = \frac{recall(N)}{N} \tag{4.2}$$

Para fins de comparação entre os algoritmos, foi calculado o gráfico de recall para cada **N** escolhido, de 1 até 20, e também calculado o gráfico de precisão por recall, esse último foi escolhido em detrimento da curva ROC por cada caso de teste possuir muitos casos não relevantes deixando o conjunto de dados muito desbalanceado, em cada caso de teste há apenas 1 verdadeiramente positivo e 1000 verdadeiramente negativos, o que causaria distorções na curva, por utilizar a quantidade de verdadeiramente negativos, que não são causados na de precisão por recall.

4.2 RESULTADOS

Nessa seção é discutido os resultados encontrados no primeiro experimento, um possível novo recomendador proposto a partir da análise do desempenho de cada recomendador no primeiro experimento e os resultados encontrados no segundo experimento, com esse novo recomendador proposto.

4.2.1 PRIMEIRO EXPERIMENTO

Como foi dito anteriormente, os testes foram feitos com 1000 casos, cada um com apenas um local verdadeiramente relevante. Para o primeiro teste, foram rodados os quatro algoritmos sobre o mesmo conjunto de testes e calculado o recall e a precisão a fim de gerar os gráficos para comparação.

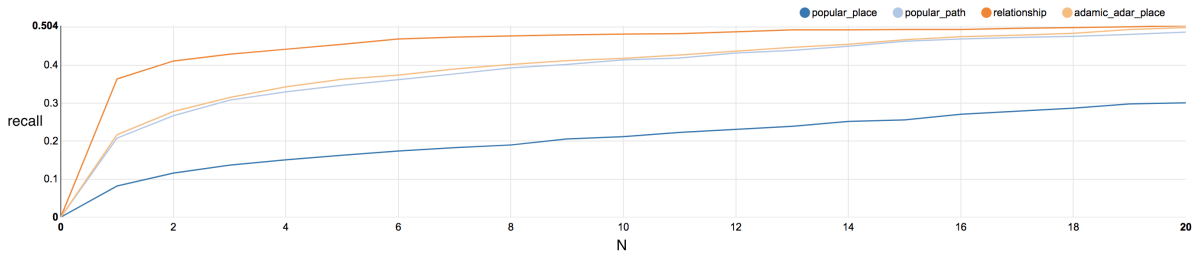

Figura 4.1 – Gráfico de recall do resultado do primeiro teste utilizando os 4 algoritmos descritos anteriormente.

Como mostra a Figura 4.1, o *Popular Place* teve o pior desempenho, se pegarmos os primeiros 20 resultados da lista de recomendação para cada caso de teste, o local relevante aparece em apenas 30% deles. Como esse algoritmo baseia sua pontuação para um local apenas considerando a quantidade de pessoas que o visitaram, há uma probabilidade alta de, ao gerar os casos de teste aleatoriamente, serem selecionados locais relevantes que não sejam muito visitados, o que afeta diretamente o desempenho desse recomendador.

Já no caso do *Popular Path* e no caso do *Adamic-Adar Places*, percebe-se que os dois tiveram desempenhos muito parecidos, com o último tendo um desempenho levemente maior, se comparado pegando os 20 primeiros resultados, o local relevante aparece em 48,7% e 49,9% das vezes, respectivamente. Esses dois algoritmos são bem parecidos, a única diferença é que o último dá um peso para as ligações ao invés de apenas contar quantos caminhos tem, o que mostra um ganho em desempenho por essa ponderação, mas nada muito relevante.

Por fim, tem o *Relationship* com o melhor desempenho entre os 4 algoritmos. Uma características interessante de notar é que até os seis primeiros resultados, ele possui um desempenho muito superior aos outros, com 46,9% dos locais relevantes já aparecendo entre eles, porém, após esse ponto, seu recall não sofre muita alteração, ficando com 50,4% considerando os 20 primeiros, um comportamento que os outros três algoritmos não tiveram. Isso pode ser explicado pelo o que foi dito no capítulo anterior, onde o subgrafo de wifis-usuários é mais esparso que o de locais-usuários, afetando a quantidade de caminhos que o algoritmo pode explorar para fazer sua recomendação.

Outra forma de visualizar esse ponto do recomendador baseado em relacionamentos é através da análise do gráfico de precisão por recall da Figura 4.2. Dá para notar que ele começa com uma precisão muito alta no seu primeiro recall, e sofre uma queda vertiginosa nos seus próximos recalls. Esse comportamento se deve ao fato de que, como a precisão é calculada a partir do recall sobre o **N** observado, como houve essa estagnação do recall, a precisão só tende a cair.

A possível explicação da estagnação do *Relationship* ser causada pela esparsidade dos dados na rede que ele se utiliza é confirmada ao olhar a cobertura de cada algoritmo na Tabela 4.1. Essa cobertura é calculada vendo quantos nós de cada caso de teste conseguem ser pontuados pelos algoritmos, por exemplo, se os algoritmos baseados em caminhos não conseguirem achar um caminho até um dos locais do caso de teste, aquele local não será pontuado.

Algoritmo	Cobertura
Popular Place	100%
Popular Path	33,68%
Adamic-Adar Places	33,68%
Relationship	3,10%

Tabela 4.1 – Cobertura de cada algoritmo nos casos de teste.

Como o *Popular Place* apenas pega todos os locais que irão sofrer a recomendação

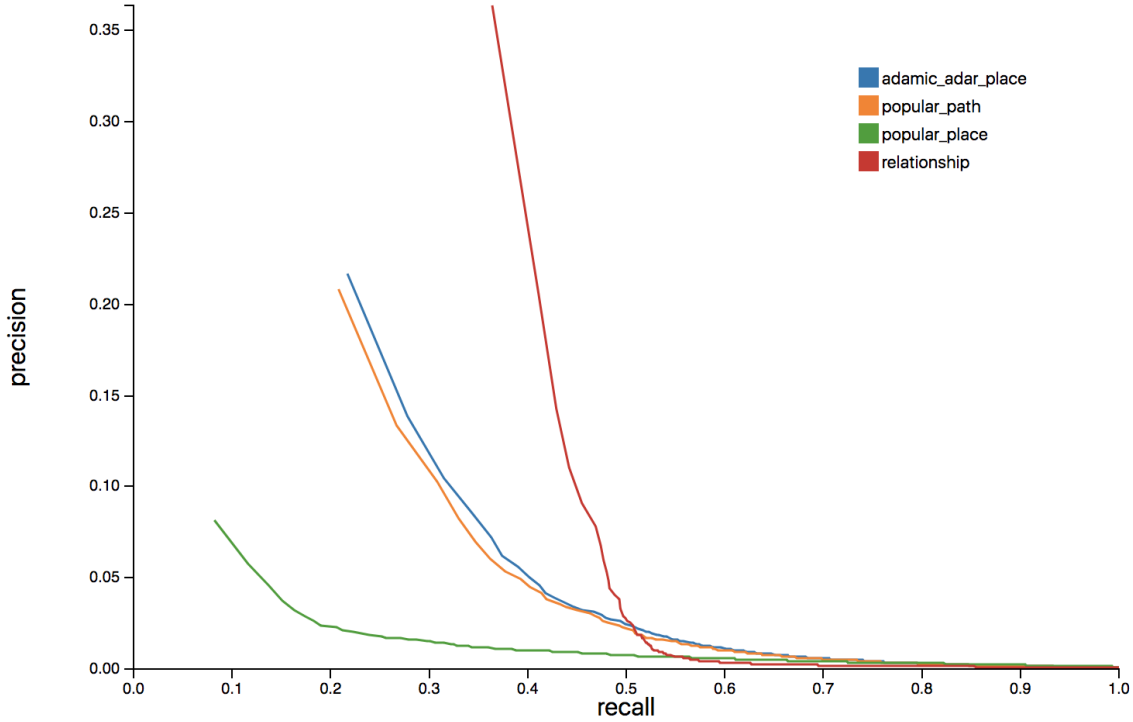

Figura 4.2 – Gráfico de precisão por recall do resultado do primeiro teste utilizando os 4 algoritmos descritos anteriormente.

e ordena baseado na quantidade de visitas, ele consegue dar uma pontuação para todos os locais. Já os outros três algoritmos dependem de encontrar caminhos entre o usuário e os locais, o que nem sempre é possível dada a esparsidade da rede. O *Popular Path* e o *Adamic-Adar Places* já sofrem com a esparsidade do subgrafo locais-usuários, porém o *Relationship* é que tem a pior cobertura por utilizar o subgrafo wifis-usuários.

Para entender melhor os dois principais algoritmos, *Relationship* e *Adamic-Adar Places*, que estão sendo testados e que utilizam ao máximo das características da rede e pelo fato de possuírem coberturas bem diferentes, foi calculado a correlação de Spearman (SPEARMAN, 1904) no resultado dos dois algoritmos para cada caso de teste e depois retirada uma média, que resultou em uma correlação de **0,2883**. Como esse resultado pode ir de **-1**, ranqueamento totalmente oposto, até **1**, ranqueamento idêntico, isso evidencia que os dois recomendadores não possuem uma relação forte entre as suas pontuações, ou seja, se um pontua bem um local, não necessariamente o outro também irá ranquear aquele local bem.

Considerando que os dois possuem essa correlação fraca, que o *Relationship* possui uma precisão alta nos menores recalls mas depois sofre pela falta de cobertura e que o *Adamic-Adar Places* possui um desempenho razoável se comparado com o algoritmo de relacionamento mas possui uma cobertura bem maior, é proposto um quinto recomendador que irá juntar os dois algoritmos para tentar se utilizar dos pontos positivos de ambos para obter um resultado melhor.

4.2.2 SEGUNDO EXPERIMENTO

Para o segundo experimento, foi proposto esse novo recomendador chamado *Mix Adamic-Adar*, que roda o *Relationship* e o *Adamic-Adar Places* e é calculado a média da pontuação de cada local dada pelos algoritmos.

Definido esse novo recomendador, foi feito um segundo experimento sobre o mesmo conjunto de casos de teste do primeiro experimento, resultando no gráfico de recall da Figura 4.3.

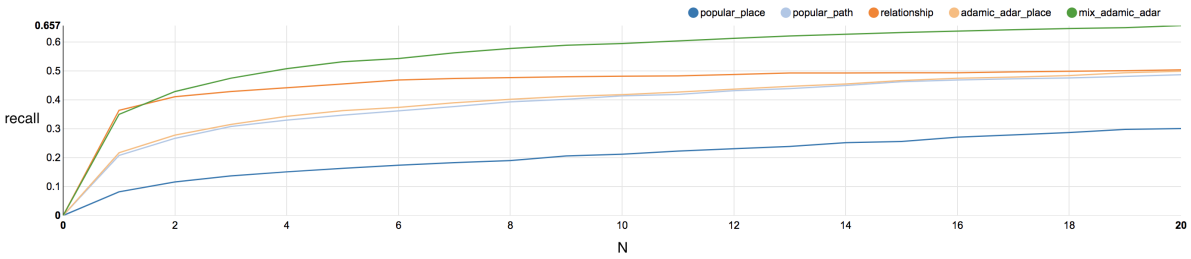

Figura 4.3 – Gráfico de recall do resultado do segundo teste utilizando os 4 algoritmos descritos anteriormente e o novo algoritmo proposto.

Como pode ser notado, esse novo algoritmo teve uma melhora considerável sobre os outros recomendadores, já pegando apenas os 4 primeiros melhores resultados de cada caso de teste, 50,8% dos locais relevantes já estão presentes, chegando até 65,7% se considerado os 20 primeiros.

Olhando para a tabela das coberturas com esse novo recomendador, disponível na Tabela 4.2, nota-se que esse algoritmo tem uma leve melhora na cobertura comparado com o *Adamic-Adar Place*, mostrando que tem alguns caminhos, apesar de poucos, que o *Relationship* cobre a mais.

Algoritmo	Cobertura
Popular Place	100%
Mix Adamic-Adar	34,05%
Popular Path	33,68%
Adamic-Adar Places	33,68%
Relationship	3,10%

Tabela 4.2 – Cobertura de cada algoritmo, com a adição do *Mix Adamic-Adar*, nos casos de teste.

Por fim, analisando o gráfico de precisão por recall da Figura 4.4, pode ser notado que, apesar do *Mix Adamic-Adar* não ter uma precisão tão alta quanto o *Relationship* no

seu primeiro recall, ele tem uma queda de precisão bem menor com os próximos recalls, resultando no melhor recomendador dentre os analisados.

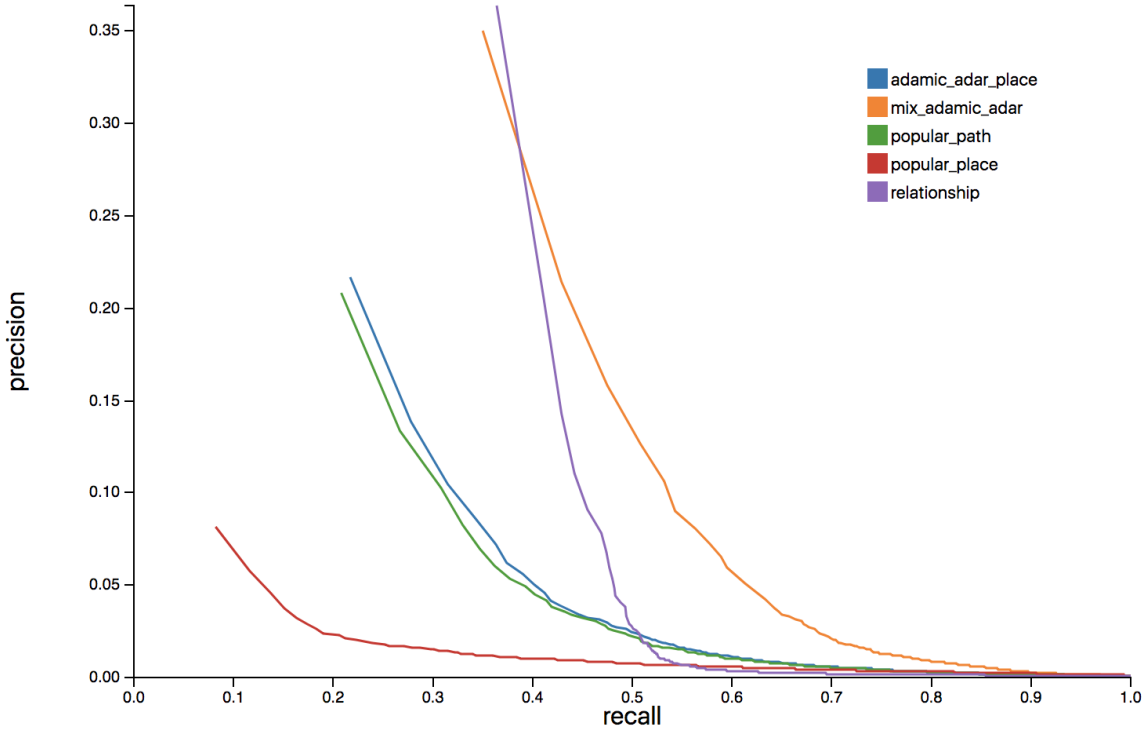

Figura 4.4 – Gráfico de precisão por recall do resultado do segundo teste utilizando os 4 algoritmos descritos anteriormente e o novo algoritmo proposto.

Com esses resultados percebe-se que o *Relationship* é um recomendador muito bom, porém tem seu desempenho diminuído pela esparsidade dos dados que usa. Ao agrupar com o *Adamic-Adar Places*, que possui uma cobertura muito maior, ele consegue ajudar a recomendar de uma maneira mais pessoal, não relacionado ao perfil das pessoas que visitam o mesmo local que um determinado usuário, mas sim relacionado ao perfil das pessoas mais próximas a ele, utilizando o conceito de homifilia, como explicado anteriormente.

5 CONCLUSÃO

Neste trabalho foi feito um estudo de técnicas de predição de link sobre redes heterogêneas formadas a partir de dados de localização e de conexão wifi. O conjunto de dados adquirido a partir do comportamento *offline* do usuário, ou seja, sem precisar de uma ação do usuário para registrar os dados de visita e conexões, permitiu pegar dados reais sobre visitas e sobre os wifis as quais ele se conectou.

O problema enfrentado pela esparsidade dos dados de conexão wifi, apesar de inicialmente os seus dados se mostrarem serem os mais precisos para recomendação, fez com que fosse necessário encontrar um jeito de juntar com outro recomendador para melhorar a cobertura.

Apesar de no final ter obtido resultados positivos, há varias possibilidades de trabalhos futuros que podem trazer uma melhora e que é interessante ser explorado.

Uma das possibilidades é extrair mais dados de conexões wifi. Os dados usados nesse trabalho foram pegos a partir das visitas, ou seja, um usuário visitou um local e, se tiver conectado a um wifi, esse dado é retornado, porém pode ter casos que a visita é registrada mas só depois dela que o usuário se conecta a algum wifi. A própria In Loco possui uma base de dados de wifi com mais dados, porém chega a ser 2 vezes o tamanho da base de visitas utilizado nesse trabalho, se tornando inviável, mas se for investido algum dinheiro para rodar as análises na nuvem, esses dados podem trazer ganhos para o recomendador de relacionamentos.

Outra possibilidade é tratar as conexões baseado em relevância para o usuário. Nesse caso, em vez de ter só a informação de se o usuário visitou aquele local ou se conectou aquela wifi, pode ser levado em conta a frequência que ele faz essa visita ou se conecta ao wifi. Com isso também pode ser possível descobrir a rotina do usuário e fazer uma recomendação baseada em dia da semana, apesar de isso causar um aumento de complexidade.

Também é possível enriquecer a rede com mais informações sobre o usuário e o local, por exemplo se o usuário é vegetariano e o local é um restaurante vegetariano, isso auxiliaria na definição do interesse do usuário no local.

Dessa forma, dá para notar que os dados de localização e relacionamento podem ser explorados considerando outros aspectos ou até mesmo outros tipos de dados, esse trabalho apenas trata de uma das diversas possibilidades que esses dados apresentam.

REFERÊNCIAS

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. *Social Networks*, v. 25, p. 211–230, Jul 2003. Disponível em: <https://doi.org/10.1016/S0378-8733(03)00009-1>. Citado na página 16.

CREMONESI, P.; KOREN, Y.; TURRIN, R. Performance of recommender algorithms on top-n recommendation tasks. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2010. (RecSys '10), p. 39–46. ISBN 978-1-60558-906-0. Disponível em: <http://doi.acm.org/10.1145/1864708.1864721>. Citado na página 25.

DAVIS, D.; LICHTENWALTER, R.; CHAWLA, N. V. Multi-relational link prediction in heterogeneous information networks. In: *International Conference on Advances in Social Networks Analysis and Mining*. [s.n.], 2011. p. 281–288. Disponível em: <https://doi.org/10.1109/ASONAM.2011.107>. Citado na página 16.

GAO, H. et al. Content-aware point of interest recommendation on location-based social networks. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2015. p. 1721–1727. Citado 2 vezes nas páginas 13 e 16.

HASAN, M. A.; ZAKI, M. J. A survey of link prediction in social networks. In: _____. *Social Network Data Analytics*. Boston, MA: Springer US, 2011. p. 243–275. ISBN 978-1-4419-8462-3. Disponível em: <https://doi.org/10.1007/978-1-4419-8462-3_9>. Citado na página 14.

KATZ, L. A new status index derived from sociometric analysis. *Psychometrika*, v. 18, p. 39–43, March 1953. Citado na página 16.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, v. 58, n. 7, p. 1019–1031, 2007. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20591>. Citado na página 15.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. In: *Annual Review of Sociology*. [s.n.], 2001. v. 27, p. 415–444. Disponível em: <https://doi.org/10.1146/annurev.soc.27.1.415>. Citado na página 14.

NEWMAN, M. E. J. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, American Physical Society, v. 64, p. 025102, Jul 2001. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.64.025102>. Citado na página 15.

OELLINGER, T.; WENNERBERG, P. O. Ontology based modeling and visualization of social networks for the web. In: *GI Jahrestagung*. [S.l.: s.n.], 2006. p. 489–497. Citado na página 12.

SCELLATO, S.; NOULAS, A.; MASCOLO, C. Exploiting place features in link prediction on location-based social networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011. (KDD '11), p. 1046–1054. ISBN 978-1-4503-0813-7. Disponível em: <http://doi.acm.org/10.1145/2020408.2020575>. Citado na página 17.

SCOTT, J. *Social Network Analysis*. [S.l.]: SAGE Publications, 2017. ISBN 9781473952119. Citado na página 11.

SPEARMAN, C. The proof and measurement of association between two things. *The American Journal of Psychology*, University of Illinois Press, v. 15, n. 1, p. 72–101, 1904. ISSN 00029556. Disponível em: <http://www.jstor.org/stable/1412159>. Citado na página 28.

WASSERMAN, S.; FAUST, K. *Social Network Analysis: Methods and Applications*. [S.l.]: Cambridge University Press, 1994. ISBN 0521382696. Citado na página 11.

ZHENG, Y. Location-based social networks: Users. In: _____. *Computing with Spatial Trajectories*. New York, NY: Springer New York, 2011. p. 243–276. ISBN 978-1-4614-1629-6. Disponível em: <https://doi.org/10.1007/978-1-4614-1629-6_8>. Citado na página 14.