UNIVERSIDADE FEDERAL DE PERNAMBUCO

BRUNO MACHADO DIAS MACENA

PREDIÇÃO DE RELACIONAMENTOS EM REDES SOCIAIS BASEADA EM LOCALIZAÇÃO

RECIFE, 2013.

BRUNO MACHADO DIAS MACENA

PREDIÇÃO DE RELACIONAMENTOS EM REDES SOCIAIS BASEADA EM LOCALIZAÇÃO

Trabalho de conclusão de curso apresentado à disciplina de Trabalho de Graduação, como requisito parcial para obtenção de grau de Bacharel em Ciência da Computação da Universidade Federal de Pernambuco.

Orientador: Ricardo Bastos Prudêncio

RECIFE, 2013.

Dedico este trabalho a minha família e em especial a minha mãe.

“Stay hungry. Stay foolish.”

Steve Jobs

AGRADECIMENTOS

Gostaria de agradecer a toda minha família pelo suporte que foi me dado durante toda a minha fase na universidade e também anteriormente nas duas tentativas para ingressar no curso. Um agradecimento em especial para a minha mãe que sempre me deu suporte quando mais precisei, me dando força para superar todos os obstáculos que enfrentei durante esse período.

Agradecer também a todos os amigos que tive a oportunidade de conhecer durante esse tempo e que também foram fundamentais nesse percurso de quatro anos e meio. Em especial a Josi, Sofia, Paulo e Angelo, que tive a sorte de conhecer no meio do curso, mas que foram essenciais nessa jornada.

Agradeço também ao professor e orientador Ricardo Prudêncio, pelo suporte oferecido nesse semestre para a conclusão deste trabalho. Estendo esse agradecimento também a todos que fazem parte do Centro de Informática e que oferecem a estrutura necessária de aprendizado para todos que estudam nesse centro.

RESUMO

O surgimento de redes sociais na internet representou uma grande mudança na forma de obter informações relevantes, que podem ser utilizadas das mais diversas formas. Nessas redes há um modelo do mundo real, com relacionamentos e assuntos de interesse dos usuários, que pode ser imensamente explorado. Uma forma de explorar esse modelo é tentar predizer relacionamentos futuros, de acordo com as informações, e utilizar isso para, por exemplo, sugerir amizades.

Mais recentemente, surgiram redes sociais como o *Foursquare* que registram dados de localização de seus usuários. Dessa forma, uma nova fonte de informação surgiu, sendo possível encontrar possíveis relações, de acordo com os locais visitados. Esse trabalho tem como objetivo estudar as técnicas e formas disponíveis de predizer possíveis relacionamentos em redes sociais na internet baseando-se na localização das pessoas.

Palavras-chave: Predição de Relacionamentos, Redes Sociais, Localização.

SUMÁRIO

1 INTRODUÇÃO

Redes sociais mudaram a forma como as pessoas se comunicam na internet. Mais do que isso, o surgimento de redes como o *MySpace*, *Facebook* e *Google+* permitiram registrar laços do mundo real que agora também existem no mundo virtual. E a partir desses registros foi possível coletar vários dados que anteriormente não eram possíveis. Em consequência disso, estudar esses relacionamentos e tentar encontrar formas de utilizar essas informações tornou-se uma importante área de pesquisa. A análise de redes sociais emergiu na tentativa de utilizar esses dados como forma de produzir ferramentas que facilitassem a vida do usuário ou até mesmo com o objetivo de utilizar essas informações comercialmente.

Uma das pesquisas que são realizadas nessa área de estudo é a predição de relacionamentos. Ou seja, de acordo com uma estrutura inicial de relacionamentos na rede prever possíveis laços futuros entre os participantes. Atualmente muitas das redes sociais utilizam-se de algoritmos para sugerir amizades aos seus usuários, de acordo com as suas amizades atuais e com outras informações que são fornecidas pelo próprio usuário no seu cadastro e durante a utilização de sua conta ao divulgar novidades e preferências em suas *timeline* (linha do tempo).

Após algum tempo, além dessas redes sociais tradicionais citadas anteriormente, surgiram redes que permitiam ao usuário compartilhar informações referentes a sua localização. É o caso do *Foursquare* e dos já extintos *Gowalla* e *Brightkite*. Esses novos dados que surgiram permitiram um novo tipo de análise em relação a predição, agora baseada na localização e nos possíveis links que poderiam surgir entre usuários de acordo com os seus locais frequentados e outros tipos de dados que essas redes fornecem. Considerando que redes como o *Foursquare* possuem mais de 40 milhões de usuários cadastrados [1], temos uma grande base de informações de onde se podem extrair várias características relevantes. Como será visto nesse trabalho, essa tipo de pesquisa baseada em localização surgiu anteriormente utilizando informações das redes móveis de celulares. No entanto, coletar os dados dessas redes nunca foi uma tarefa simples e dessa forma não era possível criar um modelo tão fiel das relações existentes entre os usuários de acordo com suas ligações.

Esse trabalho tem como objetivo então explorar os conceitos tradicionais em relação a predição de links, que normalmente levam em consideração a topologia da rede estudada, e fazer um comparativo com essa nova forma de estudo baseada na localização dos usuários. Para isso, um experimento foi realizado com o intuito de comparar técnicas de extração de características dessas redes e verificar o quanto as informações sobre os locais visitados pelas pessoas podem influenciar em relacionamentos futuros.

1.1 Estrutura do documento

Essa monografia está divida em cinco capítulos. No segundo capítulo é feita uma breve introdução sobre a Predição de Links, definindo pontos importantes para entendimento básico do tema apresentado nessa monografia. Já no terceiro capítulo é apresentada um visão da predição de relacionamentos baseada na utilização da localização dos indivíduos, apresentando os principais conceitos sobre o tema. No quarto capítulo é visto todo o detalhamento do experimento realizado com base no que foi visto anteriormente no trabalho, utilizando uma base de dados de uma rede social baseada em localização. O experimento é apresentado desde a escolha da base até as conclusões retiradas do estudo. Por fim, no quinto e último capítulo é feita a conclusão sobre o trabalho realizado e também relatado possíveis trabalhos futuros sobre o tema.

2 PREDIÇÃO DE LINKS

A estrutura de uma rede social pode ser modelada como um grafo, de forma que as relações estabelecidas pelas amizades nessa rede podem ser representadas como arestas e os nós podem representar as pessoas. As diversas áreas de estudo na Análise de Redes Sociais utilizam-se desse modelo para realizar suas pesquisas. A análise desse tipo de rede é um vasto campo de pesquisa que lida com as mais diversas técnicas e estratégias para o estudo das redes sociais [2]. E uma dessas técnicas, que possui bastante estudos sendo realizadas atualmente, é a Predição de Links.

Segundo [3], a Predição de Links é um problema de classificação que tem como objetivo detectar entre todos os possíveis pares de usuários da rede os que não estão conectados no passado e que estarão futuramente. Ou seja, dado um retrato da rede social no tempo t, o objetivo é procurar prever com a maior precisão possível quais novas arestas, que representam os relacionamentos, serão adicionadas durante o intervalo de tempo t até um dado tempo futuro t' [4].

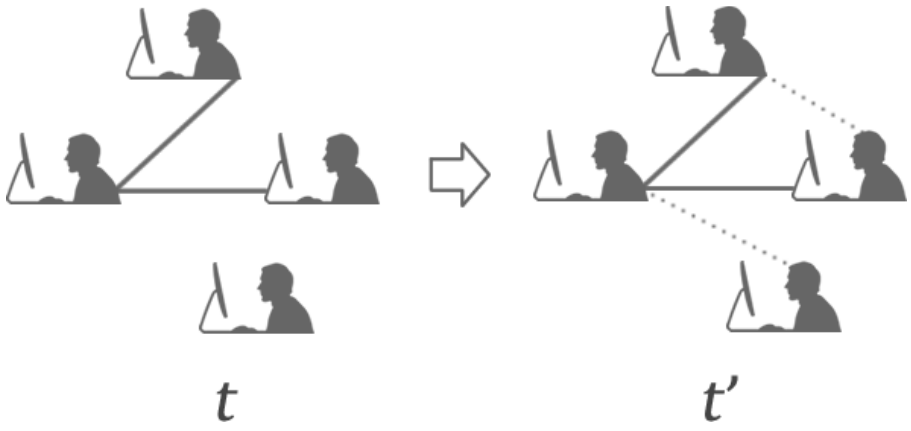


Figura 2.1: Simplificadamente é esse o objetivo da Predição de Links. Dado um conjunto de usuários (nós) no tempo *t* e suas relações (arestas), quais serão as novas relações que surgirão entre esses usuários no tempo *t'*.

Para realizar esse tipo de predição uma tarefa fundamental é extrair da rede estudada as características intrínsecas a ela. Essas características é que irão permitir o estudo da evolução da rede. São elas que servirão como medida de proximidade entre participantes de uma rede social. É possível citar, como exemplo, uma rede de coautoria entre cientistas, que inclusive costuma ser bastante utilizada

em trabalhos de Predição de Links. Nesse tipo de rede dois cientistas que estão próximos, mas que ainda não possuem uma conexão, provavelmente tem colegas em comum e consequentemente fazem parte dos mesmos círculos sociais. Com essa proximidade, há uma maior probabilidade de que esses dois cientistas trabalhem em conjunto em um projeto futuro. Uma característica como essa, extraída da topologia da rede, pode colaborar bastante no trabalho de predição. [4]

Várias outras características poderão ser exploradas, de acordo com o tipo de rede estudada, e essa escolha é fundamental para o sucesso da predição, como poderá ser visto na próxima seção.

2.1 Conjunto de características

Como foi explicado no início desse capítulo, uma tarefa importante, que servirá de base para todo o estudo da evolução da rede, é a seleção do conjunto de características que será utilizado no processo de predição.

Segundo [5], definir um conjunto de características é a parte mais difícil de qualquer algoritmo de aprendizagem de máquina. Para a Predição de Links, é necessário escolher características que representem algum tipo de proximidade ou similaridade entre o par de nós no grafo modelado da rede. No entanto, a definição dessas características pode variar de acordo com o domínio que está sendo estudado para as predições.

Um importante conjunto de características pode ser extraído a partir da topologia da rede estudada. E o mais interessante é que esse tipo de característica pode ser aplicado a qualquer domínio, diferentemente de outras características encontradas, afinal seu valor depende somente da estrutura da rede. E uma dessas características que pode ser destacada, devido a sua grande utilização, é a menor distância entre os nós. Quanto menor a distância entre dois nós, maior será a chance de que um dia eles possam estar interligados. [5]

Além desses, existem outros tantos conjuntos que podem ser criados de acordo com a necessidade de cada domínio. Só é fundamental que o conjunto escolhido represente bem o problema, como já foi dito.

No trabalho de [6], são apresentadas algumas teorias sociológicas que servem de base para a extração dessas métricas para o estudo das redes, como diz [7]. Esses aspectos sociológicos servem para demonstrar como é o comportamento das pessoas de acordo com o meio em que estão inseridas e quais são os

processos que regem a interação entre essas pessoas [6]. Algumas das teorias apresentadas por [6] estão descritas abaixo:

1) *Homofilia* – Duas pessoas que compartilham mais atributos tem maior chance de estarem conectadas futuramente que aquelas que possuem poucos atributos em comum. Ou seja, se essas duas pessoas tem atividades e interesses em comum, é mais provável que elas possuam uma relação de amizade no futuro.
2) *Raridade* – Os atributos raros, ou seja, aqueles que não são comuns, tendem a ser mais importantes comparados aqueles que são comuns. Interesses que são pouco comuns possuem normalmente um grupo menor de pessoas interessadas e essas pessoas tem maior probabilidade de se relacionar devido a essa proximidade.
3) *Influência social* – Os atributos compartilhados por uma grande porcentagem de amigos de uma determinada pessoa pode ser importante para prever possíveis links futuros para a mesma. Se muitos dos seus amigos gostam de um programa ou série de televisão em comum, é maior a chance de relacionamento com pessoas de perfil parecido.
4) *Amizades em comum* – Quanto mais amigos em comum duas pessoas compartilham, maior é a probabilidade de eles possuírem um link no futuro. Ao ter vários amigos em comum, é grande as chances de que as pessoas envolvidas frequentem locais em comum e consequentemente possuam mais chances de se conhecer, estabelecendo uma conexão real e virtual.
5) *Proximidade social* – Os amigos em potencial estão provavelmente próximos um do outro em um grafo social. Ou seja, quanto menor for a distância no grafo, maior é a chance de uma amizade existir no futuro.
6) *Conexão preferencial* – Uma pessoa tende a se relacionar com uma pessoa mais popular e não a uma pessoa com pouquíssimos amigos. Ou seja, quanto mais amigos a pessoa tiver, maior é a probabilidade de um possível novo relacionamento.

De acordo com a rede que está sendo estudada, esses comportamentos podem ser explorados e reunidos para a definição dessas medidas. Na próxima seção serão apresentadas algumas dessas métricas comuns a vários trabalhos já realizados na área.

2.2 Métricas de proximidade

Como foi falado na seção anterior, a maior parte das abordagens utilizadas na predição de links se baseia na proximidade entre os nós da rede. Para isso é necessário que se definam métricas que quantifiquem essa proximidade. Abaixo são apresentadas quatro métricas das mais utilizadas nos trabalhos de predição, que se baseiam na topologia da rede estudada [3]:

- *Vizinhos em comum (Common neighbors)* – O número de vizinhos que os nós x e y possuem em comum. Ou seja, $CN(x,y) \equiv |\Gamma(x) \cap \Gamma(y)|$, onde $\Gamma(x) \equiv \{y | y \in V, (x,y) \in E\}$ é conjunto de vizinhos do nó x.

- *Adamic-Adar* – Essa técnica é um refinamento da técnica vista anteriormente de Vizinhos em comum *(Common neighbors)*, dando um valor de peso a esses vizinhos de acordo com seus graus e não fazendo simplesmente uma contagem. Dessa forma, a contribuição dos vizinhos a técnica dos Vizinhos em comum é penalizada pelo inverso do logaritmo do seu grau.

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

- *Coeficiente de Jaccard (Jaccard's coefficient)* – Definido como o valor da intersecção dos vizinhos de dois nós, $\Gamma(x)$ e $\Gamma(y)$, dividido pela tamanho da união, caracterizando dessa forma a similaridade entre os seus conjuntos de vizinhos.

$$J(x,y) \equiv \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- *Katz* – É uma somatória de todos os caminhos possíveis do nó x até o nó y, ponderando com maior intensidade os caminhos que são mais curtos. $K(x,y) \equiv \sum_{l=1}^{\infty} \beta^l \cdot \left|paths_{x,y}^{l}\right|$, onde $paths_{x,y}^{l}$ é o conjunto de todos os caminhos que possuem tamanho l do nó x ao y. O fator β possui usualmente o valor 0.05.

Escolhida uma dessas métricas ou outras que explorem as características da rede escolhida, é necessário definir um método de avaliação que explore os valores obtidos através dessas medidas e, dessa forma, realizando a predição. Os dois métodos de avaliação estudados serão vistos na próxima seção, apresentando suas principais características de avaliação.

2.3 Métodos de avaliação

Todos os métodos utilizados para avaliar o problema da predição de links possuem um aspecto em comum. É sempre necessário, como primeiro passo, atribuir um *score* que determine o peso da conexão ao pares dos nós da rede que ainda não possuem nenhum laço e que consequentemente serão estudados. O *score* obtido é o que representa a medida de proximidade ou similaridade entre os nós. Para estabelecer quais pares de nós apresentam *scores* representativos é possível utilizar dois tipos de avaliação, a não-supervisionada e a supervisionada. [4]

Nos próximos tópicos serão apresentados os procedimentos necessários para cada um desse métodos. No experimento realizado nesse trabalho, o método escolhido para avaliação foi o não-supervisionado.

2.3.1 *Método não-supervisionado*

Nesse método, a rede incialmente é dividida em dois períodos de tempo distintos. O primeiro período é utilizado para extração de características dos pares de nós estudados. Já o segundo é utilizado para atribuição de rótulos aos pares, para indicar a ausência ou ocorrência de um relacionamento no período utilizado para predição. Esse segundo intervalo de tempo será utilizado como um teste da predição realizada de acordo com o primeiro período. [7]

No conjunto de teste, os pares de nós analisados são ordenados em ordem decrescente de acordo com o valor obtido de *score* utilizando alguma das métricas de proximidade e/ou similaridade. Ao ordenar esses nós, pode se adotar duas estratégias diferentes posteriormente. Uma delas é definir um limiar θ, definindo que todos os pares com valor maior que esse limiar estarão conectados. A outra forma é definir um parâmetro *L,* que seria a quantidade de novas arestas preditas. Não existe uma fase de treinamento para essa abordagem de avaliação. [7]

2.3.2 *Método supervisionado*

Na abordagem supervisionada, a predição de novos relacionamentos é tratada como um problema de classificação binária, no qual as duas condições para um par de nós são conectados ou não-conectados (classe positiva e negativa, respectivamente) [7].

Inicialmente a rede deve ser dividida em dois períodos distintos que não se intersectam. A primeira sub-rede, construída a partir do primeiro período de tempo, é utilizada na construção de conjunto de treinamento sendo rotulado de acordo com as métricas definidas de proximidade e similaridade. Já o segundo período será utilizado para a construção de um conjunto de teste que será utilizado para avaliar o resultado das predições realizadas no primeiro período de tempo. É necessário que os pares de nós estudados estejam presentes em ambos os períodos definidos e que não estejam conectados no período inicial. Só dessa forma será possível realizar o estudo de criação de relações futuras entre esses nós. [7]

Definidos os conjuntos de treinamento e teste, é utilizado primeiramente o conjunto de treinamento para que se treine um classificador, como, por exemplo, árvores de decisão, redes neurais e tantos outros. Feito isso, o conjunto de teste é utilizado no classificador treinado para que seja feita a avaliação do resultado. [7]

2.4 Aplicações

Métodos de predição podem ser utilizados para diferentes objetivos. Uma das possibilidades oferecidas, bastante utilizada e conhecida atualmente, é a de analisar a sua rede de amigos e sugerir novos relacionamentos baseado nesses dados. No entanto, existem outras formas de utilizar esse tipo de estudo.

Recentemente uma outra área que tem se utilizado dessa nova perspectiva é a de pesquisa em segurança, procurando monitorar possíveis redes terroristas. É

possível inferir, através dessas pesquisas, links que não podem ser explicitamente observados, permitindo desarticular conexões em células terroristas [4].

Na medicina e biologia, também existem estudos que se utilizam da predição. É possível utilizar a predição para encontrar relações e associações que existem, mas que provavelmente só poderiam ser descobertas após uma árdua e dispendiosa investigação, além de um estudo sobre uma grande variedade de agentes. [8]

2.5 Considerações finais

Como foi visto nesse capítulo, ao iniciar um projeto que tem como objetivo encontrar links em redes sociais, é necessário realizar várias escolhas inicialmente. A depender do domínio estudado, podem existir melhores formas de conduzir o processo de pesquisa. É preciso escolher um bom conjunto de características, de acordo com o domínio, além de extrair métricas que quantifiquem essas características, para que seja feita uma avaliação da rede escolhida para o estudo, com o objetivo de prever novos links no futuro.

No próximo capítulo, será estudado um domínio que vem sendo crescentemente utilizado para essas pesquisas, que é o de redes sociais que possuem informações de localização de seus usuários. Como até pouco tempo não havia nenhuma forma de rastrear esse tipo de informação, não era possível realizar nenhum tipo de pesquisa com esse objetivo. No entanto, com o advento de redes móveis de *smartphones* e o surgimento das redes sociais baseadas em localização, como *Gowalla* e *Foursquare*, surgiram novos dados que podem ser utilizados na descoberta de novos relacionamentos futuros entre as pessoas.

3 PREDIÇÃO BASEADA EM LOCALIZAÇÃO

Como já foi dito no capítulo anterior, é preciso tirar proveito das informações que são fornecidas pelo usuários para que seja possível realizar um boa predição de futuros relacionamentos nas redes sociais. São essas informações que são utilizadas como características no momento de definir qual a probabilidade de futuramente existir um relacionamento entre duas pessoas.

Com o surgimento de ferramentas como o *Foursquare,* em 2009, uma nova fonte de informação pode ser extraída das redes sociais baseadas em localização. Como esse tipo de rede tem como princípio fundamental o compartilhamento da localização, através do comumente chamado *check-in* (uma atividade online que diz aos seus amigos quando e onde você está através de uma rede social [9]*),* entre amigos, é possível aproveitar esse novo aspecto, não utilizado em outros serviços sociais, na inferência de novos links nesse tipo de rede.

Como diz [10], todas as redes sociais baseadas em localização possuem o que se intitula de um *framework* “3+1”, ou seja, três camadas e uma *timeline*, como é mostrado na figura 3.1.

A camada geográfica possui todo o histórico de *check-ins* realizados pelos usuários, a social contém informações sobre as amizades e a de conteúdo é uma camada que consiste de *feedbacks* ou dicas sobre os mais diversos locais que foram visitados pelos usuários. Todas essas camadas compartilham uma única *timeline*, onde todas essas informações costumam ser exibidas [7]. É a camada geográfica que diferencia as redes sociais baseadas em localização das mais tradicionais. As informações de todas as outras camadas já podiam ser utilizadas nas pesquisas de predição anteriores.

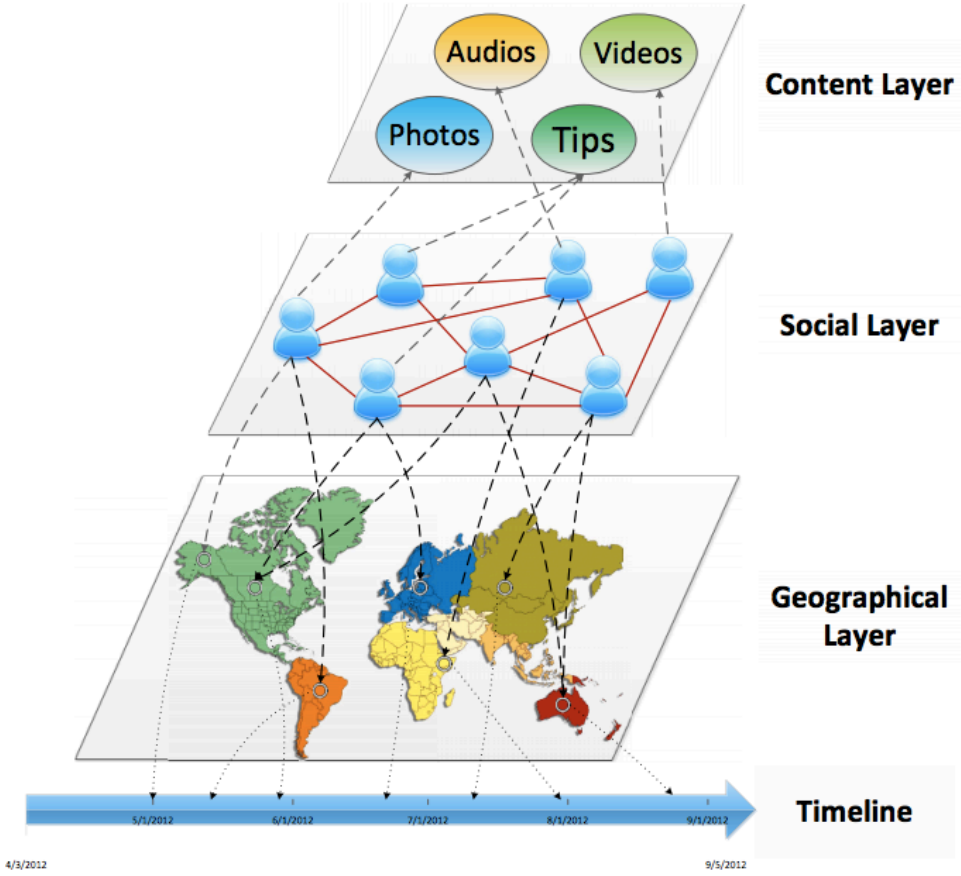


Figura 3.1: Representação em camadas de uma rede social baseada em localização como, por exemplo, *Foursquare* e *Gowalla*. A novidade desse tipo de rede é a camada geográfica que representa os pontos de localização mais comuns dos usuários cadastrados no serviço.

No entanto, a utilização das redes sociais não foi a primeira tentativa de utilizar informações de localização na tentativa de prever possíveis futuros relacionamentos entre as pessoas. Como será visto na próxima seção, esse tipo de trabalho teve início com o surgimento de redes de telefonia celular e a popularização dos smartphones.

3.1 O início da utilização da localização

Com já foi dito, com o advento das redes sociais foi possível utilizar diversas novas informações dos usuários disponíveis no espaço social virtual. No entanto, essa não foi a primeira tentativa de utilizar a localização das pessoas para a predição de links futuros.

Com o grande crescimento de uso de *smartphones* nos últimos anos, a realidade do mundo mudou bastante e permitiu as pesquisas de predição a

utilização de informações sobre a mobilidade de cada indivíduo. Esse é o ponto principal do trabalho [3], que foi um dos primeiros a utilizar o enfoque de localização nesse tipo de estudo. Como diz [3], os registros das ligações móveis coletados pelas operadoras de telefonia móvel forneceram várias informações sobre as trajetórias de cada indivíduo e as suas relações sociais, mantendo o controle de cada telefonema realizado entre duas partes e a localização no espaço e tempo da pessoa que realizou a ligação.

Como houve uma grande adoção por boa parte da população desses aparelhos móveis, isso implica na captura de dados de uma grande parcela de um país, por exemplo. Foram esses dados que tornaram possível o início das pesquisas que tentavam mensurar a importância do fator localização na predição desses novos links. A partir desses dados foi possível, pela primeira vez, ter informações sobre a rotina diária de várias pessoas, além do tempo gasto em cada um dos locais visitados. [3]

3.2 Entropia de local *(Place Entropy)*

O conceito de entropia de um local pode ajudar na previsão de que este é um bom lugar para que novas relações sejam estabelecidas. Será que um local bastante visitado é um lugar propício para que se estabeleçam novos laços de amizade ou isso ocorre mais em locais pouco frequentados?

Segundo [11], quando um usuário realiza regularmente *check-ins* em um local onde outros poucos usuários costumam fazer o mesmo, é sinal de que aquele lugar possui um significado especial para ele. É o caso, por exemplo, de casa, academia e local de trabalho, que costumam ser frequentados rotineiramente e por relativamente poucas pessoas.

Já quando um local possui *check-ins* realizados por vários usuários é bastante provável que seja um lugar público, sem qualquer importância para os participantes da rede social que realizaram o *check-in*, como, por exemplo, locais turísticos, aeroportos, estações de metrô e assim por diante [11]. Esse tipo de local costuma ser pouco relevante na criação de novos laços de amizade.

Dessa forma, uma medida do quanto um local pode promover novos relacionamentos deve levar em conta tanto o número de usuários que realizaram *check-in* no local como também a quantidade total de *check-ins* realizados. [10]

Algumas das métricas que serão vistas, levam em consideração a entropia do local para avaliar o quanto esse lugar é propício para o estabelecimento de novos amizades e, dessa forma, é possível prever alguns relacionamentos baseado nessa característica de determinado local.

3.3 Medidas de colocalização

Baseado no que foi visto anteriormente, são definidas medidas (métricas) responsáveis por quantificar a importância da colocalização na criação de novos relacionamentos. Nesta seção serão apresentadas algumas formas de realizar essa tarefa.

Quando dois usuários fazem *check-in* nos mesmos lugares eles podem ter muitas chances de estar em contato um com o outro e posteriormente é possível que se crie uma nova conexão entre eles. Por conta disso, são definidas duas medidas no trabalho [11] que tentam capturar essa característica. A primeira é a $common\_p$ que representa o número de locais em comum que duas pessoas visitaram. A outra, $overlap\_p$, leva em consideração a mesma característica, mas se diferencia por ser representada por uma fração. Nessa fração é também considerada o total de locais visitados por dois usuários.

$$common\_p \equiv \left|\Phi_i \cap \Phi_j\right|$$

$$overlap\_p \equiv \frac{\left|\Phi_i \cap \Phi_j\right|}{\left|\Phi_i \cup \Phi_j\right|}$$

Outra medida definida em [10], leva em consideração o número de *check-ins* feito por ambos usuários e é denominada w_common_p. É possível também medir a similaridade de cossenos dos vetores de *check-in* de cada usuário, definida como $w\_overlap\_p$.

$$w\_common\_p \equiv \vec{c_i}\,\vec{c_j}$$

$$w\_overlap\_p \equiv \vec{c_i}\,\vec{c_j} \Big/ \sqrt{\vec{c_i}^2\,\vec{c_j}^2}$$

Como já foi visto anteriormente, uma outra forma de fazer essa medição de proximidade, baseada em colocalização, é utilizar a entropia dos locais que dois usuários já visitaram. A primeira medida estabelecida por [10] é a min_ent que é definida como a menor entropia de local de todos os locais visitados em comum por dois usuários. Já a medida aa_ent é estabelecida como a soma dos inversos de cada valor de entropia dos locais. Essa é uma medida que possui como inspiração o *score* de similaridade Adamic-Adar, que já foi mencionado anteriormente nesse trabalho.

$$\min\ \_ent \equiv \min\left(E_k : m_k \in \Phi_i \cap \Phi_j\right)$$

$$aa\_ent \equiv \sum_{m_k \in \Phi_i \cap \Phi_j} \frac{1}{E_k}$$

Além dessas, mais duas medidas são definidas considerando o número de *check-ins* realizados. Para esse cálculo é considerado que um local compartilhado possui maior relevância, caso possua poucos *check-ins*.

$$aa\_p \equiv \sum_{m_k \in \Phi_i \cap \Phi_j} \frac{1}{\log C_k^P}$$

$$\min\_p \equiv \min\left(C_k^P : m_k \in \Phi_i \cap \Phi_j\right)$$

As métricas definidas acima foram definidas mais recentemente, pois como já foi falado, redes sociais baseadas em localização são serviços relativamente novos e as informações utilizadas para esses cálculos só puderam ser coletados com o surgimento delas. No entanto, também já foi visto que as pesquisas na área de predição de links baseando-se na localização, surgiram com o advento das redes de telefonia celular. Com base nessas informações das redes de celular, no trabalho [3] foram definidas algumas métricas. Essas medidas tem como fundamental característica a tentativa de capturar a distância física entre dois usuários da rede. Algumas delas são vistas logo abaixo:

- *Distância* – Considerando que $ML(x) \equiv argmax_{l \in Loc} PV(x,l)$ representa a localização mais provável de um usuário x, sendo Loc o conjunto de locais já visitados por x, e $PV(x,l) \equiv \sum_{i=1}^{n(x)} \delta(l, L_i(x))/n(x)$ a probabilidade dele visitar um determinado local l, podemos definir a métrica abaixo, que representa a distância física entre os locais mais frequentados pelos usuários x e y.

$$d(x,y) \equiv dist(ML(x), ML(y))$$

- *Taxa de colocalização espacial (Spatial Co-Location Rate)* – Essa medida representa a probabilidade que dois usuários x e y visitem o mesmo local, mas não necessariamente no mesmo horário. Considerando que a probabilidade de visitas para dois usuários é independente, essa métrica é definida da seguinte forma:

$$SCoL(x,y) \equiv \sum_{l \in Loc} PV(x,l) \times PV(y,l)$$

- *Similaridade de cosseno espacial (Spatial Cosine Similarity)* – A similaridade de cossenos das trajetórias dos usuários x e y, tentando detectar o quão similar são as frequências de visitas desses usuários. Essa medida é definida como o cosseno do ângulo entre os vetores de números de visitas em cada local para os usuários x e y.

$$SCos(x,y) \equiv \sum_{l \in Loc} \frac{PV(x,l) \times PV(y,l)}{\|PV(x,l)\| \times \|PV(y,l)\|}$$

Algumas outras métricas são definidas no trabalho [3], mas essas podem ser consideradas as principais.

A partir das métricas definidas nessa seção é possível definir um ranking dos possíveis nós futuros que podem surgir em uma rede. Como foi visto, de acordo

com a rede estudada podem ser aplicadas diferentes medidas baseadas na colocalização. Aqui foram vistas algumas métricas utilizadas nos estudos de redes de celulares e nas mais recentes redes sociais da internet, como o *Foursquare*.

3.4 Trabalhos relacionados

Existem outros trabalhos que se utilizam da localização sob um aspecto diferente. Uma das possibilidades é utilizar os dados não para predizer novos relacionamentos, mas sim novos locais que poderiam ser visitados pelo usuário, de acordo com os que foram visitados anteriormente. A partir desse estudo é possível se beneficiar dessa informação em áreas como marketing mobile, planejamento de tráfego ou até mesmo na assistência em desastres [10]. Existem dois pontos de pesquisas principais nessa área de estudo. Uma delas é conseguir definir a localização da casa de um determinado usuário e a outra é predizer a localização desse usuário em tempo real.

Outra forma de se utilizar dessas informações é na criação de sistemas de recomendação. Com as informações fornecidas é possível recomendar locais, *tags* e amigos.

Como se pode perceber, existem diversas formas de se utilizar desses novos dados, que podem gerar vários benefícios para vários serviços e seus respectivos usuários.

4 EXPERIMENTAÇÃO

Nesse capítulo serão detalhados todas as etapas do experimento realizado nesse projeto, como a metodologia utilizada e os resultados obtidos. Será visto como foi construída toda a experimentação, desde a escolha do conjunto de dados a ser estudado até as conclusões obtidas ao final do processo. Esse experimento tem como objetivo realizar um comparativo entre a eficácia de métricas de similaridade, comumente utilizadas em vários domínios, com as novas métricas baseadas na localização dos usuários. Portanto algumas delas citadas anteriormente nesse trabalho serão usadas para verificar o quão precisas são as predições utilizando a base de dados escolhida previamente.

4.1 Foursquare

O *Foursquare* é um aplicativo que permite aos usuários compartilhar com seus amigos sua localização, através de *check-ins*. Além disso, o serviço oferece recomendações personalizadas e dicas baseadas na atual localização do usuário.

Atualmente a rede social possui mais de 40 milhões de usuários cadastrados, com mais de 4.5 bilhões de *check-ins* já realizados [1]. E a cada dia são realizados milhões de check-ins por todo mundo.

Devido a todos esses motivos, o *Foursquare* foi a rede social escolhida para se realizar os experimentos desse trabalho. Por possuir a comunidade mais ativa atualmente das redes sociais baseadas em localização o *Foursquare* parece ser a melhor escolha para se realizar um experimento desse tipo.

4.2 Conjunto de dados

Para o experimento desse trabalho será utilizado um conjunto de dados do *Foursquare* que possui *check-ins* realizados por 18.107 usuários. Esses dados foram coletados no período de Março de 2010 a Janeiro de 2011. A base utilizada possui a rede de relacionamentos de cada usuário, como também os locais onde ele realizou *check-in*, com o respectivo horário. No total foram registrados 2.073.740 *check-ins* durante o período de tempo utilizado para o registro dos dados. Além disso, existem 115.104 relacionamentos entre esses usuários cujos dados serão utilizados para esse experimento. [9]

4.3 Preparação de dados

Como o método utilizado no experimento é o não-supervisionado, o primeiro passo é definir os períodos que serão utilizados para a extração de características e rotulação dos nós, respectivamente. Para determinar esses conjuntos foi utilizada o *RapidMiner,* ferramenta conhecida *open-source* utilizada na mineração de dados. A divisão dos 115.104 relacionamentos existentes na base foi realizada da seguinte forma: 70% desses relacionamentos fazem parte do primeiro conjunto, no qual será feita a extração de características, e os 30% restantes caracterizam o conjunto de teste, que será rotulado de acordo com a predição feita. Portanto, são 80.902 exemplos para o período de extração e 34.672 para o de teste. Esses dois conjuntos foram gerados randomicamente a partir do arquivo que possui todos os relacionamentos da base.

4.4 Métricas escolhidas

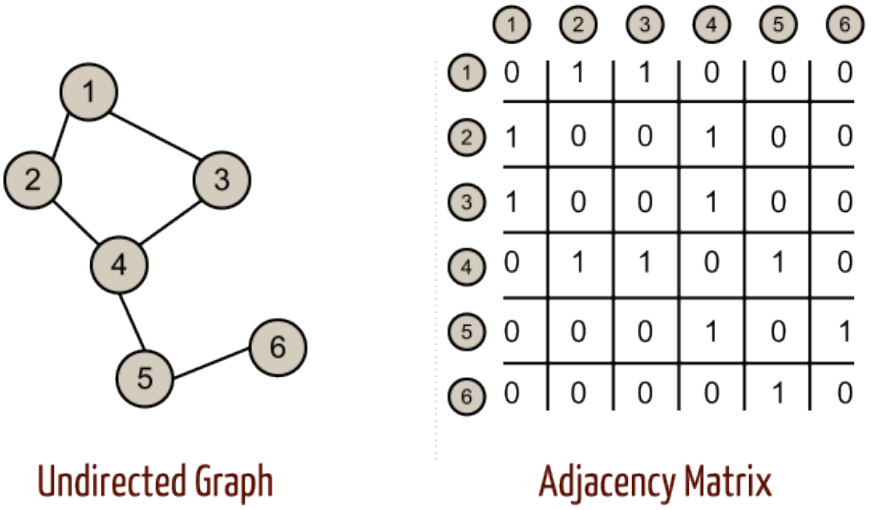
Para realizar um comparativo entre as métricas baseadas na topologia da rede, que são comumente utilizadas em vários trabalhos, e as baseadas em informações de localização, foram escolhidas duas métricas representantes de cada grupo. Dessa forma, serão computados os *scores* dos relacionamentos de acordo com as seguintes métricas:

- *Common Neighbors* - $CN(x,y) \equiv |\Gamma(x) \cap \Gamma(y)|$
- *Adamic-Adar* - $AA(x,y) \equiv \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|}$
- *Common Places (common_p)* - $common\_p \equiv |\Phi_i \cap \Phi_j|$
- *Overlap Places (overlap_p)* - $overlap\_p \equiv \frac{|\Phi_i \cap \Phi_j|}{|\Phi_i \cup \Phi_j|}$

O cálculo dessas métricas foi realizado através de um código em Java e os *scores* de todos os 327.747.875 novos relacionamentos possíveis foi calculado. Calculados todos os *scores* o passo seguinte é ranquear em ordem decrescente todos os relacionamentos de acordo com o valor de *score* obtido. Nessa tarefa também foi utilizada a ferramenta *RapidMiner*.

Para o cálculo dos *scores* de cada um dos possíveis links que podem existir no futuro, foram utilizadas matrizes de adjacência para a representação do grafo. As

informações são retiradas de uma lista de nós adjacentes e para esses nós que já possuem um link é dado o valor 1 na matriz. Todos os outros que não possuem ainda um relacionamento recebem o valor 0. A representação em formato de matriz de adjacência pode ser vista na figura 4.1.



	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

Figura 4.1: Forma de representação de um grafo através de uma matriz de adjacência. Para os cálculos dos scores no experimento essa foi a representação utilizada.

4.5 Resultados

Após realizar todas as etapas do experimento, podemos comparar os resultados obtidos de acordo com a classificação feita, baseada nos *scores* calculados. Para fazer essa comparação foram escolhidos os 500 primeiros relacionamentos do ranking que foram preditos. Desses foram selecionados 10%, 25%, 50% e 100%, sequencialmente, para avaliar a precisão que se conseguia nas predições utilizando cada uma das métricas escolhidas. O resultado pode ser visto no gráfico abaixo:

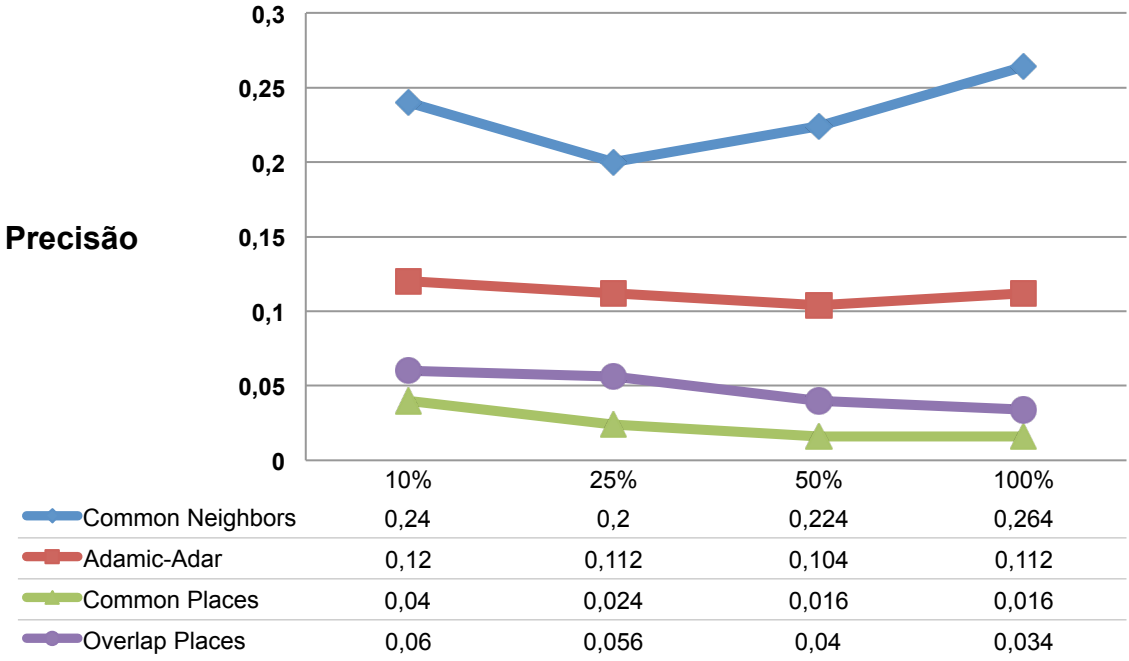


Figura 4.1: Gráfico que faz um comparativo baseado na precisão de cada uma das métricas escolhidas para o experimento, de acordo com uma certa porcentagem de links escolhidos no ranking estabelecido.

Como é possível verificar na figura 4.1, os melhores resultados foram obtidos com a métrica de Vizinhos em comum (Common Neighbors). Levando em consideração os 500 relacionamentos, ou seja, 100% dos links que foram utilizados para comparação com a base de teste, se obteve uma precisão de 26,4%. A segunda melhor métrica Adamic-Adar também é baseada em similaridade, mas obteve como melhor resultado uma precisão de 12% considerando somente 10% dos primeiros 500 links ranqueados. Por último vieram as métricas baseadas em localização com 4% (Common Places) e 6% (Overlap Places) ao se verificar também 10% do ranking de relacionamentos preditos.

É possível perceber então que para essa base de dados as métricas baseadas em localização não obtiveram um desempenho superior ou semelhante as métricas mais tradicionais. No entanto, isso não invalida a importância desse tipo de informação para obter uma maior precisão nas predições. A quantidade limitada de informações sobre os *check-ins* na base pode ter influência no resultado obtido. Além disso, intervalos de tempo maiores, que capturassem mais relacionamentos da rede estudada, poderiam influenciar no resultado final. Alguns dos links que foram

descobertos podem não existir nessa janela de tempo estudada, mas possuem boas chances de existirem futuramente.

5 CONCLUSÃO

Esse trabalho procurou apresentar conceitos mais antigos da área de pesquisa relacionada a predição de relacionamentos, que tem como base principalmente a topologia da rede estudada e apresentar uma nova perspectiva que surgiu nos últimos anos com o aparecimento de redes sociais que agora apresentam informações de localização dos seus usuários. Com base nisso, foram apresentadas algumas métricas, propostas em alguns trabalhos, que procuram tirar proveito dessas novas informações para identificar os possíveis links futuros que podem existir entre as pessoas nessas redes.

O uso de informações de localização para predição de relacionamentos é algo relativamente recente e consequentemente ainda é necessário a realização de mais pesquisas para que se possa definir métricas que extraiam mais dessas redes, podendo dessa forma obter maior precisão nas predições realizadas. Métricas que se baseiam, por exemplo, na topologia da rede estudada ainda possuem melhores resultados como pode ser visto no experimento realizado.

No entanto, foi possível constatar que essa camada adicional com dados de localização realmente representa uma nova forma de obter informações relevantes. O conceito de entropia de um local, por exemplo, que não foi explorado no experimento devido a falta de informações na base, pode servir como uma importante característica a ser estudada, permitindo calcular *scores* mais relevantes na tentativa de predição.

5.1 Trabalhos futuros

Um possível trabalho futuro seria construir uma base de dados que possua ao menos dois intervalos de tempo e que possua uma grande quantidade de relacionamentos nesses dois períodos. Além disso, seria importante também coletar informações sobre os locais visitados para que se possa utilizar métricas que explorem, por exemplo, a entropia de uma local. Uma base maior com mais informações também permitiria extrair novas características que talvez fossem relevantes e pudessem ser utilizadas em novas métricas.

Outra possibilidade, já explorada por alguns outros trabalhos, é mesclar características da topologia da rede e de localização para realizar melhores predições.

6 REFERÊNCIAS BIBLIOGRÁFICAS

1. SOBRE o Foursquare. Disponivel em: <pt.foursquare.com/about>. Acesso em: 10 Setembro 2013.
2. DE SÁ, H. R.; PRUDÊNCIO, R. B. C. Supervised link prediction in weighted networks. In: _____ **Neural Networks (IJCNN), The 2011 International Joint Conference on**. [S.l.]: IEEE, 2011. p. 2281-2288.
3. BARABÁSI, A.-L. et al. Human Mobility, Social Ties, and Link Prediction. In: _____ **Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.]: ACM, 2011. p. 1100-1108.
4. LIBEN-NOWELL, D.; KLEINBERG, J. The Link-Prediction Problem for Social Networks. **Journal of the American society for information science and technology**, p. 1019-1031, 2007.
5. AL HASAN, M. et al. Link Prediction using Supervised Learning. In: _____ **SDM'06:** Workshop on Link Analysis, Counter-terrorism and Security. [S.l.]: [s.n.], 2006.
6. YIN, Z. et al. A unified framework for link recommendation using random walks. In: _____ **Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on**. [S.l.]: IEEE, 2010. p. 152-159.
8. LICHTENWALTER, R. N.; LUSSIER, J. T.; CHAWLA, N. V. New perspectives and methods in link prediction. In: _____ **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.]: ACM, 2010. p. 243-252.
7. SOARES, P. R. D. S. **Predição de Links em Redes Sociais com Uso de Informações Temporais**. UFPE. Recife. 2012.
9. GAO, H.; TANG, J.; LIU, H. Exploring Social-Historical Ties on Location-Based Social Networks. In: _____ **ICWSM**. [S.l.]: [s.n.], 2012.
10. GAO, H.; LIU, H. Data Analysis on Location-Based Social Netwoks. In: CHIN, A.; ZHANG, D. **Mobile Social Networking:** An Innovative Approach. [S.l.]: [s.n.].
11. 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. [S.l.]: ACM, 2006. p. 1046-1054.