Processo Sistemático Baseado em Métricas Não-Dicotômicas para Avaliação de Predição de *Links* em Redes de Coautoria

Elisandra Aparecida Alves da Silva[1], **Marco Túlio Carvalho de Andrade**[2]

[1] Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP)
Av. Francisco Samuel Lucchesi Filho, 770. Penha. 12929-600
Bragança Paulista – SP – Brasil

[2]Depto. Eng. Computação e Sistemas Digitais (PCS)
Escola Politécnica Universidade de São Paulo
Av Prof. Luciano Gualberto, 158 travessa 3 – 05508-900
São Paulo – SP – Brasil

elisandra@ifsp.edu.br, mtcandrade@usp.br

Abstract. *Link prediction is an important research line in the Social Network Analysis context, as predicting the evolution of such nets is a useful mechanism to improve and encourage communication among users. In co-authorship networks, it can be used for recommending users with common research interests. This paper presents a systematic process based on non-dichotomic metrics for evaluation of link prediction in co-authorship networks considering the definition of methods for the following tasks: data selection and new link determination. Fuzzy sensor based on node attributes is adopted for data selection. Fuzzy compositions are used to predict new link weights between two authors, adopting not only attributes nodes, but also the combination of attributes of other observed links. The link weight called "relation quality" is obtained by using structural features of the social network. The AUC is used for results evaluation.*

Resumo. *Predição de Links é uma área de pesquisa importante no contexto de Análise de Redes Sociais tendo em vista que predizer sua evolução é um mecanismo útil para melhorar e propiciar a comunicação entre usuários. Nas redes de coautoria isso pode ser utilizado para recomendação de usuários com interesses de pesquisa comuns. Este artigo apresenta um processo sistemático baseado em métricas não-dicotômicas para avaliação de predição de links em redes de coautoria, sendo considerada a definição de métodos para as seguintes tarefas identificadas: seleção de dados e determinação de novos links. Para seleção de dados definiu-se um sensor fuzzy baseado em atributos dos nós. O uso de composições fuzzy foi considerado para determinação de novos links "ponderados" entre dois autores, adotando-se não apenas atributos dos nós, mas também a combinação de atributos de outros links observados. O link ponderado é denominado "qualidade da relação" e é obtido pelo uso de propriedades estruturais da rede. Para avaliação dos resultados foi adotada a AUC obtida a partir da curva ROC.*

1. Introdução

Atualmente muitas bases de dados são descritas como uma coleção de objetos inter-relacionados por *links*[1]. As redes formadas por tais objetos podem ser homogêneas, nas quais há um único tipo de objeto e de *link*, ou heterogêneas, nas quais objetos e *links* podem ser de múltiplos tipos. Como exemplo de rede homogênea, tem-se a rede de coautoria abordada neste trabalho e de rede heterogênea a *World Wide Web*.

Para [Liben-Nowell and Kleinberg 2007] as redes sociais são objetos bastante dinâmicos, que se alteram rapidamente a partir da ocorrência de novas interações na estrutura social. Dessa forma, entender os mecanismos que regem a evolução dessas redes é uma questão fundamental ainda não bem compreendida, que poderia melhorar e propiciar a comunicação entre seus integrantes. A investigação desses mecanismos, a análise das propriedades básicas e de características estruturais recorrentes é em parte motivada pela disponibilidade de grandes conjuntos de dados ([Watts and Strogatz 1998], [Watts 1999], [Grossman 2002], [Newman 2002], [Adamic and Adar 2001], [Newman 2003]). Exemplos bastante difundidos de redes sociais formadas para o estabelecimento de relações pessoais e profissionais são: Facebook, Orkut, Twitter, Linkedin, entre outras. Tais redes são heterogêneas, pois seus objetos e *links* podem ser de múltiplos tipos.

Numa rede social virtual, o usuário compartilha informações com outros parceiros que possuem interesses similares, o que lhe permite buscar informações eficientemente. Por esse motivo, as redes sociais representam uma nova forma de acesso à informação, que ganha cada vez mais força. E se comparada a *Web* no aspecto de sobrecarga de informação apresenta algumas vantagens, por reduzir o espaço de busca. Predição de *Links* é uma importante área de pesquisa no contexto de Análise de Redes Sociais tendo em vista que predizer a evolução de tais redes é um mecanismo útil para melhorar e propiciar a comunicação entre usuários.

A avaliação de Predição de *Links* realizada em diferentes trabalhos não considera a adoção de um processo sistemático, que pode ser útil na identificação de tarefas, bem como, na definição de métodos para cada uma das tarefas identificadas.

O principal objetivo de Predição de *Links* é determinar a existência de um *link* entre duas entidades usando atributos de objetos e de outros *links*. Predição de *Links* é útil em diferentes domínios de aplicação, tais como: detecção de ligações não-observadas em redes de terrorismo, redes de interação de proteínas, predição de colaborações entre cientistas e predição de hiperlinks *Web*.

Neste contexto, este artigo apresenta um processo sistemático baseado em métricas não-dicotômicas para avaliação de Predição de *Links* em redes de coautoria englobando a definição de métodos e técnicas adequadas para as tarefas identificadas. Na seção 2, apresenta-se o processo proposto para avaliação e predição de *links* e na seção 3 os experimentos realizados. Finalmente, as conclusões e referências são apresentadas.

2. Processo Proposto

O processo proposto engloba as tarefas apresentadas na Figura 1. Para seleção de dados foram adotados atributos dos nós (autores) e uma representação qualitativa, que é

[1] O termo *link* é utilizado neste artigo para representar ligações/relações entre nós de diferentes tipos, tais como: autores em redes de coautoria, usuários em redes sociais, páginas *Web* e vértices de grafos.

mais próxima da linguagem natural. As novas ligações foram determinadas a partir da composição de atributos dos nós e de outros *links* observados, verificando a importância das ligações mais recentes. E para avaliação dos resultados foi adotada a *Area Under Curve* (AUC) obtida a partir da curva ROC.

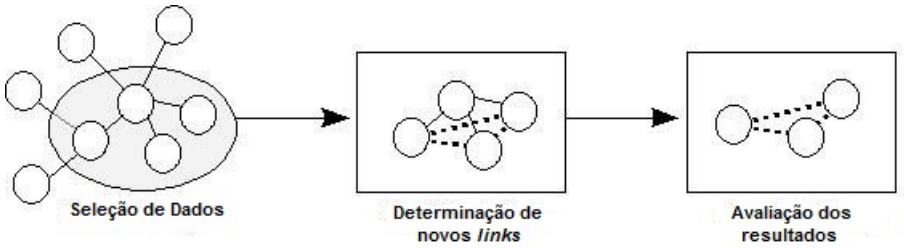

Figura 1. Processo de avaliação de predição de *links*

Os métodos propostos para cada tarefa são descritos nas próximas seções. Primeiramente, apresenta-se o método de Seleção de Dados.

2.1. Método Proposto para Seleção de Dados

[Liben-Nowell and Kleinberg 2007] observaram que a avaliação de métodos de Predição de *Links* utiliza alguns parâmetros para determinação do conjunto $Core$ usado como foco da avaliação, dentre eles, o número de ligações (ou publicações em redes de co-autoria) é bastante considerado e informações como assuntos abordados e áreas de publicações também são explorados, quando disponíveis. O sensor *fuzzy* definido para seleção dos dados engloba duas variáveis de entrada: $NúmeroDePublicações$ e $NúmeroDeCoautores$ e uma variável de saída que determina a escolha do autor. Em outros tipos de redes sociais tais variáveis representariam o número de encontros e o número de vizinhos do participante.

A Figura 2 mostra os conjuntos *fuzzy* das variáveis linguísticas $NúmeroDePublicações$, $NúmeroDeCoautores$ e $FatorDeSeleção$.

Entradas

$NúmeroDePublicações$ representa as publicações realizadas pelo coautor nos períodos de treinamento e teste (Universo de discurso: -1 a 10; Valores linguísticos: baixo, alto).

$NúmeroDeCoautores$ representa o total de coautores do autor nos períodos de treinamento e de teste (Universo de discurso: -1 a 40; Valores linguísticos: baixo, alto).

Saída

$FatorDeSeleção$ determina se autor faz parte do conjunto $Core$ (Universo de discurso: 1 a 10; Valores linguísticos: baixo, médio, alto).

Adotou-se a seguinte **base de regras *fuzzy*** no formato *if-then*:

if NúmeroDeCoautores é baixo AND NúmeroDePublicações é baixo THEN FatorDeSeleção é baixo

if NúmeroDeCoautores é baixo AND NúmeroDePublicações é alto THEN FatorDeSeleção é alto

if NúmeroDeCoautores é alto AND NúmeroDePublicações é baixo THEN FatorDeSeleção é baixo

if NúmeroDeCoautores é alto AND NúmeroDePublicações é alto THEN FatorDeSeleção é médio

A partir das regras apresentadas, verifica-se que o fator de seleção é alto quando

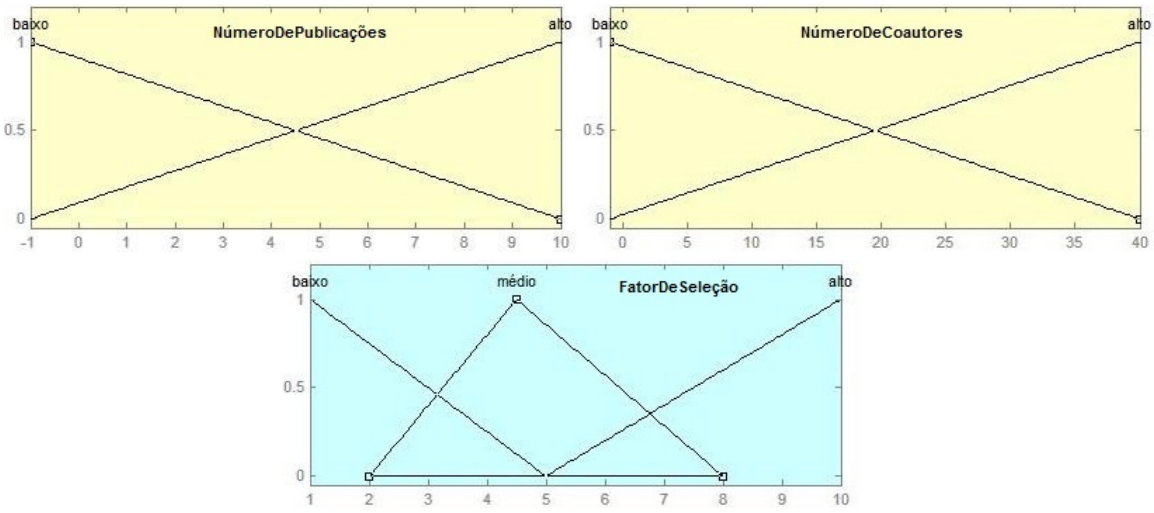

Figura 2. Funções de pertinência das variáveis linguísticas *NúmeroDePublicações*, *NúmeroDeCoautores* e *FatorDeSeleção*

o autor realizou muitas publicações com poucos coautores. Este conhecimento é explorado de forma intuitiva a partir da adoção de uma linguagem bem próxima da natural, ou seja, a partir das regras e variáveis *fuzzy* definidas pode-se explorar melhor conhecimento e não apenas trabalhar com limites. Ao final, o resultado é defuzzificado e caso o $FatorDeSeleção$ seja maior do que 5, o nó é adicionado ao $Core$.

Nas próximas seções apresentam-se os métodos propostos para determinação de novos *links* a partir de composições *fuzzy*.

3. Métodos Propostos para Determinação de Novos *Links*

Neste trabalho considera-se o peso do *link* entre dois autores x e y como a "qualidade da relação". Essa medida é obtida pela utilização de atributos de autores e ligações da rede de coautoria.

Dois métodos baseados na teoria de conjuntos *fuzzy* são propostos. Ambos adotam o uso de composições *fuzzy* para determinar novos *links* entre dois autores e aplicam a qualidade da relação para determinar o peso de um *link*. No Método *Fuzzy1* é utilizado um modelo *fuzzy* para determinar a qualidade da relação e no Método *Fuzzy*2 foi definida uma abordagem tradicional para qualidade da relação.

Os métodos consideram que a qualidade da relação entre dois autores é maior nas seguintes situações: (1) quando dois autores têm um grande número de publicações, sendo que, a valorização ou não das ligações mais recentes é analisada; (2) quando a média de coautores dos autores na relação é baixa. Entretanto, os coautores comuns influenciam a relação positivamente. Para estabelecer esses critérios experimentos adicionais foram realizados e são apresentados em [Silva and Andrade 2011].

Essa medida representa a qualidade da relação entre dois autores. Utilizando-se a qualidade da relação e composições *fuzzy* determina-se o peso de um novo *link*.

A seguir apresenta-se o método *Fuzzy1*.

3.1. Método *Fuzzy1*

As variáveis de entrada definidas neste método são: $NúmeroDePublicações$, $MédiaDeCoautores$ e $TempoDeRelação$.

NúmeroDePublicações é o número de publicações escritas em parceria por A e B. Ao valorizar as publicações mais recentes consideram-se diferentes pesos de acordo com o ano da publicação. E o valor que uma publicação adiciona ao número total de publicações é obtido como segue:

$$AnoDePublicação - InícioDoPeríodoDeTreinamento$$

Dessa forma, o primeiro ano do período de treinamento não é relevante para a medida e a soma total é o número de publicações em parceria por A e B. Como os métodos são avaliados considerando a valorização e a não valorização das ligações mais recentes, caso as publicações mais recentes não sejam valorizadas, cada nova publicação adiciona 1 ao total de publicações.

MédiaDeCoautores é a média de coautores de A e B, mas os coautores comuns não são considerados. $\Gamma(A)$ é o conjunto de coautores de A e $\Gamma(B)$ é o conjunto de coautores de B. Esse valor é obtido como segue:

$$Co = \frac{|\Gamma(A)| + |\Gamma(B)|}{2} - |\Gamma(A) \cap \Gamma(B)|$$

TempoDeRelação é a diferença entre o último ano de treinamento e o ano da publicação mais antiga realizada em parceria.

As regras usadas são apresentadas a seguir:

if MédiaDeCoautores é baixa AND NúmeroDePublicações é baixo AND TempoDeRelação é baixo THEN QualidadeDaRelação é média

if MédiaDeCoautores é baixa AND NúmeroDePublicações é baixo AND TempoDeRelação é alto THEN QualidadeDaRelação é baixa

if MédiaDeCoautores é baixa AND NúmeroDePublicações é alto THEN QualidadeDaRelação é alta

if MédiaDeCoautores é alta AND NúmeroDePublicações é baixo THEN QualidadeDaRelação é baixa

if MédiaDeCoautores é alta AND NúmeroDePublicações é alto THEN QualidadeDaRelação é média

O $TempoDeRelação$ é importante nos casos em que a $MédiaDeCoautores$ é baixa e o $NúmeroDePublicações$ é baixo. Nestes casos, o tempo de relação é usado para determinar se a qualidade é baixa ou média.

3.2. Método *Fuzzy2*

A segunda abordagem baseia-se na seguinte métrica proposta para Qualidade da Relação. O valor que uma publicação adiciona ao total de número de publicação, considerando-se maior peso para as publicações mais recentes, é obtido como segue:

$$AnoDePublicação - InícioDoPeríodoDeTreinamento$$

$$Co = \frac{|\Gamma(A)| + |\Gamma(B)|}{2} - |\Gamma(A) \cap \Gamma(B)|$$

$$QualidadeDaRelação = \frac{p(A,B)}{Co}$$

sendo que $p(A, B)$ é o número de publicações feitas em parceria por A e B. Caso não se considere a valorização das publicações mais recentes este valor é incrementado a cada publicação.

4. Método Adotado para Avaliação dos Resultados

Segundo [Prati et al. 2008], a análise *Receiver Operating Characteristic* (ROC) é um método gráfico que permite avaliar sistemas de diagnóstico e/ou predição favorecendo a visualização da multidimensionalidade do problema.

Os gráficos ROC foram propostos inicialmente para analisar a qualidade de transmissão de sinais [Egan 1975]. E atualmente introduzidos como uma ferramenta poderosa para avaliação de classificadores nas áreas de Aprendizagem de Máquina e Mineração de Dados [Bradley 1997], [Spackman 1989].

A curva ROC é uma representação bidimensional do desempenho de um classificador, baseada na probabilidade de detecção (taxa de verdadeiros positivos) no eixo y, e de falsos alarmes (taxa de falsos positivos) no eixo x. A taxa de verdadeiros positivos também é denominada $recall$ ou $sensitivity$. Para gerar a curva ROC essas taxas são determinadas em vários pontos de corte e não apenas em um único limiar, o que permite uma análise independente do limiar.

Para avaliar o desempenho de um classificador adotando-se a curva ROC, deve-se verificar sua distância da diagonal principal, sendo que quanto mais distante melhor é o desempenho para o domínio considerado. No melhor caso, a curva deve ser convexa e sempre crescente [Prati et al. 2008]. Portanto, quando é necessário comparar o desempenho de dois ou mais classificadores, a curva que mais se aproxima do ponto $(0, 1)$ é a de melhor desempenho. Entretanto, podem ocorrer intersecções, e nestes casos, cada um dos classificadores tem uma faixa operacional em que é melhor do que o outro [Prati et al. 2008].

Uma estratégia comum utilizada para comparar classificadores é reduzir o desempenho a um único valor escalar que possa representá-la adequadamente. Para isso, é comum calcular a *Area Under Curve* (AUC) [Bradley 1997], [Hanley and Mcneil 1982]. Esta área é uma porção da área do quadrado de lado 1, portanto, seu valor está entre 0 e 1 [Prati et al. 2008]. Classificadores aleatórios produzem a linha diagonal, que tem uma área de 0.5, de forma que, nenhum classificador deveria ter uma AUC menor que 0.5 [Fawcett 2006].

Segundo [Fawcett 2006], a AUC tem uma propriedade estatística importante: é numericamente igual à probabilidade de dados dois exemplos de classes distintas, o exemplo positivo seja ordenado primeiramente que o exemplo negativo. E para [Acar et al. 2009], que adotoram a AUC como métrica para avaliação de desempenho de predição de *links*, é uma métrica robusta em domínios com classes desbalanceadas, como ocorre nas redes de coautoria.

5. Experimentos

Adotando-se o processo proposto, diferentes experimentos foram realizados permitindo comparar os métodos definidos a métodos já conhecidos para cada uma das tarefas identificadas. Portanto, na tarefa de seleção de dados usados como foco da avaliação adota-se

o sensor *fuzzy* proposto e um método tradicional. Para a tarefa de determinação de novos *links* dois métodos baseados em composições *fuzzy* (*Fuzzy1* e *Fuzzy2*) são comparados aos métodos tradicionais que também utilizam propriedades estruturais da rede. Para verificar se as ligações mais recentes melhoram o resultado da classificação, foram realizados experimentos valorizando as ligações mais recentes e não diferenciando os pesos das ligações. Na tarefa de avaliação dos resultados foi adotada a AUC obtida a partir da curva ROC.

5.1. Seleção de Dados

A tarefa de Seleção de Dados é responsável por determinar o conjunto $Core$. Este conjunto representa um subconjunto da rede utilizada para determinação de novos *links*. De forma geral, o método de determinação de novos *links* gera uma lista de possíveis *links*, que se pretende verificar na rede num período futuro. Para avaliar os resultados do método, o foco da avaliação são os nós que pertencem ao conjunto $Core$.

Apresentam-se na Tabela 1 os períodos utilizados e informações sobre a base de dados em cada período. E_{old} representa as ligações no período de treinamento e E_{new} as ligações novas no período de teste. Na Tabela 2 observam-se informações adicionais sobre a base DBLP nos períodos, sendo número de autores e publicações.

A base DBLP (Digital Bibliography & Library Project) contém dados de publicações da área de Ciência da Computação e tem sido utilizada em diferentes trabalhos na área de Predição de *Links* [Hasan et al. 2006], [Wang et al. 2007], [Acar et al. 2009],[Scripps et al. 2008].

A base bibliográfica DBLP da University of Trier contém mais de 1.15 milhões de registros e detalhes de publicações de conferências relacionadas às áreas de Mineração de Dados, Banco de Dados, Aprendizado de Máquina e outras. DBLP é pública e está no formato no XML [Trier 2009]. Para extração dos dados foi implementado um *parser* Java que coleta as informações necessárias para aplicação dos métodos. Todos os métodos propostos e analisados foram implementados em Java.

A base DBLP representa a rede de coautoria escolhida para aplicação do processo de avaliação de Predição de *Links*. A primeira tarefa realizada na avaliação é a Seleção de Dados, apresentada a seguir.

Período	Treinamento	Teste	$\|E_{old}\|$	$\|E_{new}\|$
1	1999-2004	2005-2007	663530	745629
2	2000-2006	2007-2009	1057817	523844

Tabela 1. $|E_{old}|$ e $|E_{new}|$ nos períodos de treinamento e teste

Período	Autores (Treinamento)	Public. (Treinamento)	Autores (Teste)	Public. (Teste)
1	303615	377143	330324	338712
2	416808	551713	277499	245252

Tabela 2. Números de autores e publicações nos períodos

A seguir apresentam-se os dados obtidos com o uso do método tradicional de geração do $Core$.

5.2. Método Tradicional de Geração do $Core$

O método tradicional de geração do conjunto $Core$ citado por [Liben-Nowell and Kleinberg 2007] adota os parâmetros $k_{training}$ e k_{test}. O conjunto $Core$ é formado por nós que possuem no mínimo $k_{training}$ ligações em $G[t_0, t'_0]$(grafo do período de treinamento) e no mínimo k_{test} ligações em $G[t_1, t'_1]$ (grafo do período de teste).

Para realização dos experimentos utilizando o método tradicional foram considerados $k_{training} > 1$ e $k_{test} > 1$, ou seja, o conjunto $Core$ é formado por nós que tenham no mínimo duas (2) publicações no período de treinamento e no mínimo duas (2) publicações no período de teste.

Na Tabela 3 apresenta-se o número de ligações em E^*_{new}, que são as ligações no intervalo de teste que fazem parte do conjunto $Core$, nos dois períodos.

Período	Treinamento	Teste	$\lvert E^*_{new}\rvert$
1	1999-2004	2005-2007	114894
2	2000-2006	2007-2009	99676

Tabela 3. $\lvert E^*_{new}\rvert$ usando método tradicional de geração do $Core$

Os dados obtidos pelo uso do Sensor *Fuzzy* são apresentados a seguir.

5.3. Sensor *Fuzzy* para Geraçao do $Core$

Neste trabalho, um sensor *fuzzy* é definido para geração do conjunto $Core$. O sensor é formado pelas variáveis de entrada $NúmeroDePublicações$ e $NúmeroDeCoautores$ e por uma variável de saída que determina se o nó pertencerá ou não ao conjunto $Core$.

Na Tabela 4 apresenta-se o número de ligações em E^*_{new} nos dois períodos considerando-se a aplicação do sensor *fuzzy*.

Período	Treinamento	Teste	$\lvert E^*_{new}\rvert$
1	1999-2004	2005-2007	151840
2	2000-2006	2007-2009	127903

Tabela 4. $\lvert E^*_{new}\rvert$ usando o sensor *fuzzy*

Tendo-se selecionado os dados do conjunto $Core$, a próxima tarefa é a determinação de novos *links*.

5.4. Determinaçao de Novos *Links*

Considerando-se os dois métodos de seleção de dados e os períodos apresentados, foram utilizados diferentes métodos para predição de novos *links*, sendo esses: Vizinhos Comuns, *Preferential Attachment*, Adamic/Adar e os métodos *Fuzzy1* e *Fuzzy2*, que utilizam composições *fuzzy*, que são comparados a métodos baseados em propriedades semelhantes, ou seja, no uso de propriedades estruturais da rede. Uma das diferenças básicas está na adoção de composições *fuzzy* que permite combinar atributos de *links* observados e não apenas atributos dos nós para determinação do peso, ou $score$, de um novo *link*.

A abordagem básica desses métodos é a classificação de todos os pares de nós a partir de medidas de proximidade do grafo. O peso do *link* denominado $score(x, y)$ é

atribuído a cada par de nós x e y, e então uma lista é gerada em ordem decrescente de $score$. Considerando-se o nó x, $\Gamma(x)$ denota o conjunto de vizinhos de x em G_{collab} (grafo que representa a rede). Os vizinhos de x são os nós que estão diretamente conectados a x.

Dessa forma, esses métodos podem ser vistos como a computação da medida de proximidade entre os nós x e y, relacionada à topologia da rede e, em geral, são derivados da Teoria dos Grafos e da Análise de Redes Sociais. Segundo [Liben-Nowell and Kleinberg 2007], esses métodos precisam ser adaptados para aplicação em diferentes contextos.

Muitos métodos baseiam-se na idéia de que quanto maior o número de vizinhos comuns entre dois objetos maior a chance de existir um *link* entre x e y. [Davidsen et al. 2002] e [Jin et al. 2001] propuseram modelos abstratos para crescimento da rede usando esta idéia. Eles apresentam a aplicação mais direta de Vizinhos Comuns para Predição de *Links*, sendo que [Newman 2001] usou essa medida no contexto de redes de colaboração. Dessa forma, o peso do *link* entre os usuários x e y é obtido pela intersecção dos conjuntos de vizinhos de x e y, ou seja, representa os vizinhos comuns desses usuários, como segue:

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

[Adamic and Adar 2001] usaram a idéia de proximidade para verificar a similaridade entre páginas pessoais da *Web*. Eles assumem que vizinhos comuns com graus mais baixos, ou seja, menor número de vizinhos, são mais relevantes, da seguinte maneira:

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

Outro método, denominado *Preferential Attachment*, assume que a probabilidade de um novo *link* entre x e y é proporcional ao número de *links* dos vizinhos. Essa medida é obtida como segue [Barabási et al. 2002]:

$$score(x, y) = |\Gamma(x)| \times |\Gamma(y)|$$

Os métodos de Predição de *Links* apresentados são baseados em propriedades estruturais da rede e não consideram os pesos das ligações entre os usuários. [Murata and Moriyasu 2008] propuseram algumas adaptações baseadas em medidas de proximidade que foram aplicadas em redes sociais *online*. Como as informações pessoais dos usuários não estão geralmente disponíveis nessas redes, somente propriedades estruturais foram utilizadas.

Os métodos *Fuzzy1* e *Fuzzy2*, ambos baseados em composições *fuzzy*, foram analisados considerando-se a valorização e a não valorização das ligações mais recentes. Os métodos Vizinhos Comuns, Adamic/Adar, e *Preferential Attachment* não se baseiam no uso de ligações, desta forma, a valorização das ligações recentes não tem impacto nesses métodos.

A seguir, apresenta-se a avaliação dos resultados adotando-se a AUC.

5.5. Avaliação dos Resultados

Para avaliação foram utilizados dois períodos para considerar diferentes números de autores e publicações. Apresentam-se na Tabela 5 os valores das AUCs obtidas para o período 1 e na Tabela 6 os valores das AUCs obtidas para o período 2.

$Core$	VizCom	Prefer	Adamic	Fuz1NãoVal	Fuz1Val	Fuz2NãoVal	Fuz2Val
Tradicional	0.5870	0.5373	0.5923	0.6134	0.6007	0.5979	0.5977
Fuzzy	0.5851	0.5391	0.5942	0.6207	0.6045	0.6025	0.5969

Tabela 5. AUCs obtidas no período 1

Observa-se pela Tabela 5, que usando o método tradicional de geração do $Core$ no período 1, o *Preferential Attachment* apresentou o pior desempenho e o método *Fuzzy1* não valorizando as ligações mais recentes obteve o melhor. A valorização das ligações recentes piorou os resultados das predições para os métodos *Fuzzy* propostos. O sensor *fuzzy* para geração do $Core$ não apresentou diferenças em relação ao método tradicional de geração do $Core$ na determinação do pior e do melhor método, e a maior variação entre as AUCs do método tradicional para o sensor *fuzzy* foi para o método *Fuzzy1* e *Fuzzy2* não valorizando as ligações recentes.

$Core$	VizCom	Prefer	Adamic	Fuz1NãoVal	Fuz1Val	Fuz2NãoVal	Fuz2Val
Tradicional	0.5751	0.5695	0.6075	0.6161	0.6058	0.6186	0.6128
Fuzzy	0.5720	0.5719	0.6071	0.6244	0.5948	0.6123	0.6045

Tabela 6. AUCs obtidas no período 2

A partir da Tabela 6, observa-se que usando o método tradicional de geração do $Core$ no período 2, o *Preferential Attachment* apresentou o pior desempenho e o método *Fuzzy2* não valorizando as ligações mais recentes obteve o melhor, mas bem próximo do obtida para o método *Fuzzy1* também não valorizando as ligações mais recentes. A valorização das ligações recentes piorou os resultados das predições para os métodos *Fuzzy* propostos também no período 2. O sensor *fuzzy* para geração do $Core$ não apresentou diferenças em relação ao método tradicional de geração do $Core$ na determinação do pior método, mas usando o sensor *fuzzy* o melhor método foi o *Fuzzy1* não valorizando as ligações recentes. A maior variação entre as AUCs do método tradicional para o sensor *fuzzy* foi para o método *Fuzzy1* valorizando e não valorizando as ligações recentes e *Fuzzy2* valorizando as ligações recentes.

Analisando-se o método tradicional e o sensor *fuzzy* verifica-se que ambos resultaram em desempenhos bastante próximos. O uso do sensor *fuzzy*, entretanto, permite selecionar objetos que podem ser desconsiderados no método tradicional que utiliza uma representação dicotômica e, em geral, apenas uma variável. O sensor *fuzzy* permite utilizar outras variáveis e adotar uma representação mais próxima da natural.

De forma geral, analisando-se os resultados obtidos para os métodos *Fuzzy1* e *Fuzzy2* verifica-se que ambos tiveram melhor desempenho do que os demais nos dois períodos utilizando-se o método tradicional ou o sensor *fuzzy* para geração do conjunto $Core$.

6. Conclusões

Os resultados mostram que a aplicação do sensor *fuzzy* para determinação do conjunto $Core$ gerou resultados bem próximos do método tradicional adotado na literatura por [Liben-Nowell and Kleinberg 2007]. A principal vantagem está na forma bastante intuitiva de expressar o conhecimento.

O uso de composições *fuzzy* permitiu considerar atributos de nós e outros *links* na determinação do peso de um novo *link*. O modelo *fuzzy* para determinar a Qualidade da Relação é interessante, pois permite que o conhecimento do especialista no domínio seja aproveitado na definição das variáveis, visto que algumas características são inerentes ao tipo de rede social. O modelo proposto, entretanto, pode ser adequado a diferentes domínios considerando-se a definição de novas variáveis ou mesmo adaptações das variáveis propostas.

Para definição da métrica e das variáveis do modelo *fuzzy* considerou-se que a Qualidade da Relação entre dois autores é maior nas seguintes situações: (1) quando dois autores possuem muitas publicações em parceria; (2) quando a média de coautores dos autores da relação é baixa, desconsiderando-se os vizinhos comuns, visto que estes influenciam positivamente a relação.

Pode-se concluir pelos resultados apresentados que o método *Preferential Attachment* apresentou o pior desempenho nos dois períodos. Segundo [Murata and Moriyasu 2008], o método não é apropriado para redes com graus de distribuição uniformes. Adamic/Adar apresentou melhores resultados do que Vizinhos Comuns para os dois períodos e é considerado um método estável que apresenta bons desempenhos em diferentes domínios. Os métodos *Fuzzy* apresentaram melhores resultados do que os demais nos dois períodos e representam uma forma intuitiva de trabalhar com as variáveis do domínio.

Referências

Acar, E., Dunlavy, D. M., and Kolda, T. G. (2009). Link prediction on evolving data using matrix and tensor factorizations. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 262–269, Washington, DC, USA. IEEE Computer Society.

Adamic, L. A. and Adar, E. (2001). Friends and neighbors on the web. *SOCIAL NETWORKS*, 25:211–230.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Davidsen, J., Ebel, H., and Bornholdt, S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(12):128701.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. Academic Press, New York, USA.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

Grossman, J. W. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, 158:201–212.

Hanley, J. A. and Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hasan, M., Chaoji, V., Salem, S., and Zaki, M. J. (2006). Link prediction using supervised learning.

Jin, E. M., Girvan, M., and Newman, M. E. J. (2001). The structure of growing social networks. *Physical Review E*, 64(4):046132.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

Murata, T. and Moriyasu, S. (2008). Link prediction based on structural properties of online social networks. *New Generation Comput.*, 26(3):245–257.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA*, 98(2):404–409.

Newman, M. E. J. (2002). The structure and function of networks. *Computer Physics Communications*, 147:40–45.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Prati, R. C., Batista, G. E. A. P. A., and Monard, M. C. (2008). Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, 6(2):215–222.

Scripps, J., Tan, P.-N., Chen, F., and Esfahanian, A.-H. (2008). A matrix alignment approach for link prediction. *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4.

Silva, E. A. A. and Andrade, M. T. C. (2011). *Proposta de um Processo Sistemático Baseado em Métricas Não-Dicotômicas para Avaliação de Predição de* Links *em Redes de Coautoria*. PhD thesis, Escola Politécnica da Universidade de São Paulo.

Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the 6th Int Workshop on Machine Learning*, pages 160–163.

Trier, U. (2009). Digital bibliography & library project (dblp).

Wang, C., Satuluri, V., and Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 322–331, Washington, DC, USA. IEEE Computer Society.

Watts, D. J. (1999). *Small Worlds*. Princeton University Press, New Jersey.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.