Universidade Federal do Estado do Rio de Janeiro

Centro de Ciências Exatas e Tecnologia

Programa de Pós-Graduação em Informática

PREDIÇÃO SEMÂNTICA DE UNLINKS EM REDES EVOLUTIVAS UTILIZANDO LÓGICA DE DESCRIÇÃO PROBABILÍSTICA

Marcius Armada de Oliveira

Orientadores

Kate Cerqueira Revoredo

José Eduardo Ochoa Luna

Rio de Janeiro, RJ - Brasil

Setembro de 2014

PREDIÇÃO SEMÂNTICA DE UNLINKS EM REDES EVOLUTIVAS UTILIZANDO LÓGICA DE DESCRIÇÃO PROBABILÍSTICA

Marcius Armada de Oliveira

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Kate Cerqueira Revoredo, D.Sc. - UNIRIO

Fernanda Araujo Baião, D.Sc. - UNIRIO

Jonice de Oliveira Sampaio, D.Sc. - UFRJ

Rio de Janeiro, RJ - Brasil

Setembro de 2014

Armada, Marcius
A727 Predição Semântica de Unlinks em Redes Evolutivas Utilizando Lógica de Descrição Probabilística / Marcius Armada de Oliveira, 2014.
83 f. ; 30 cm

Orientadora: Kate Cerqueira Revoredo.
Coorientador: José Eduardo Ochoa Luna.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

1. Ontologia (Informática). 2. Rede social on line. 3. Predição semântica de Unlink. 4. Predição de Link. 5. Rede evolutiva. 6. Lógica de descrição probabilística. I. Revoredo, Kate Cerqueira. II. Luna, José Eduardo Ochoa. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD-006.312

À minha família.

AGRADECIMENTOS

Aos meus pais que foram sempre o exemplo pra tudo na minha vida, de caráter, de trabalho e de família. Se eu sou o homem que sou hoje em dia é porque vocês foram os grandes orientadores da minha vida.

À minha irmã sempre amiga que sempre torceu pelo meu sucesso, confiante da minha capacidade para alcançá-lo. Tenho que agradece-la não só pelo apoio mas pelo acréscimo à nossa família do Pedro, seu marido, e agradecer aos dois pela felicidade de me fazerem tio pela primeira vez.

Aos primos Cristina e Marcelo, irmãos que a vida me deu, obrigado por estarem sempre presentes nos momentos mais importantes.

À Dona Dulce e à Dona Luiza que, ainda durante o meu bacharelado e antes mesmo do meu casamento, me acolheram com tanto carinho na sua casa e nas suas vidas e contribuíram com muito incentivo nesta minha jornada.

Aos meus "entefilhos", Carol e Lucas, que fizeram de mim seu "paidrasto". Aprendo todos os dias um pouco mais o significado destes sentimentos: de querer bem, proteger, orientar, ser o melhor exemplo possível e me orgulhar das suas conquistas.

Aos amigos do mestrado pelas conversas, ideias e palavras de incentivo que trocamos nesta empreitada que compartilhamos. Espero repetir nossas "bancadas etílicas" em breve.

Ao parceiro de estudos e pesquisa, da graduação, Cleomar que foi o responsável pelo "empurrão" inicial que resultou na inscrição e aprovação para o Mestrado na Unirio. Sem esse incentivo eu não estaria aqui agradecendo e comemorando mais

essa vitória. Obrigado também a Cristina que, naquela época, nos aturou durante os longos dias e noites de trabalho.

Aos amigos da Fundação COPPETEC que seguraram mais de uma vez as "pontas" para que eu pudesse estudar, concluir um artigo ou viajar para algum evento. Sem esse apoio e suporte nada disso teria sido possível.

Aos professores do PPGI que impulsionam com dedicação o sucesso desta instituição de ensino e principalmente aqueles que de qualquer forma contribuíram para minha formação.

À banca pelas suas palavras e orientações. À Jonice que eu tive oportunidade de conhecer no SBSI 2103, viajamos lado a lado no avião, quem diria viria fazer parte da minha banca. Do pouco que tive oportunidade de conversar se mostrou uma entusiasta e grande pesquisadora da área. À Fernanda que acompanhou toda essas jornada sempre bem próxima. Participou de algumas bancas de seminário e contribuiu desde sempre com suas observações para o desenvolvimento deste trabalho.

À Kate, pela oportunidade quando solicitei a troca de orientador, pois percebi uma maior afinidade com a sua linha de pesquisa, não imaginei a sorte e alegria de me deparar com essa grande orientadora que você é. Pela sua paciência diante das minhas dificuldades e pela orientação calma e constante. Você pra mim é um exemplo de professora e do que a academia tem de melhor. Além de tudo isso uma amiga! Agradeço também ao José que, apesar de estar longe, coorientou este trabalho e sempre que necessário contribui pontualmente com todo o seu conhecimento.

À Kátia, minha esposa, meu amor, razão para tantas mudanças positivas na minha vida. Todo meu esforço pra fazer nossa vida cada vez mais feliz é pouco para agradecer todo o seu carinho e dedicação. Te amo muito!

Obrigado!

ARMADA, Marcius. **Predição Semântica de Unlinks em Redes Evolutivas Utilizando Lógica de Descrição Probabilística**. UNIRIO, 2014. 70 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Recentemente, a predição de links entre indivíduos em redes sociais tem recebido bastante atenção. Entretanto, para entender completamente e predizer como uma rede social evolui através do tempo, além de predizer os links que serão adicionados a rede, também é necessário predizer os que serão removidos. A predição de unlinks é a tarefa de predizer se um par de nós conectados irá se desligar. Neste trabalho, propomos uma abordagem semântica, que utiliza a informação a respeito do domínio em questão, para a predição de unlinks através de uma ontologia probabilística. Avaliamos empiricamente nossa proposta comparando-a aos métodos estruturais mais utilizados na predição de links, mas utilizando-os inversamente para predizer unlinks, e ao estado da arte em métodos para a predição de unlinks utilizando duas redes de colaboração científicas: Lattes e DBLP. Os resultados mostram melhora significativa na detecção de unlinks quando nossa proposta é considerada.

Palavras-chave: Predição de Unlink, Predição semântica de Unlink, Predição de Link, Rede Social, Rede Evolutiva, Lógica de Descrição Probabilística, Ontologia Probabilística.

ABSTRACT

Recently, prediction of new links between two individuals in social networks has gain a lot of attention. However, to fully understand and predict how the network evolves through time, in addition to predicting the links that are added, ending relationships also need to be predicted. The unlink prediction is the task to predict if a connected pair of nodes will disconnect. In this work, we propose a semantic approach that uses information about the domain of discourse to predict unlinks through a probabilistic ontology. We empirically evaluated our approach comparing it with the most used graph-based link prediction methods, but inversaly using it to predict unlinks, and with the state of the art unlink methods using two scientific collaboration networks: Lattes and DBLP. The results show significant improvement on detecting unlink when considering our proposal.

Keywords: Unlink Prediction, Semantic Unlink Prediction, Link Prediction, Social Network, Evolving Network, Probabilistic Description Logic, Probabilistic Ontology.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Este capítulo fornece uma visão geral da dissertação, bem como a motivação para a pesquisa da tarefa de predição de unlinks. A proposta de solução para o problema é brevemente descrita, apresentando as linhas de ação que norteiam esta dissertação. O objetivo, a hipótese e a metodologia usada para testá-la também são apresentadas.

Muitos domínios podem ser representados por redes onde nós simbolizam objetos ou indivíduos e links denotam relações ou interações entre esses objetos. Estas redes apresentam um comportamento dinâmico, onde nós e links podem aparecer e desaparecer com o tempo.

Neste cenário, predizer um possível link na rede, isto é, a ocorrência futura de um novo relacionamento entre dois nós é um problema interessante que vem recebendo significativa atenção [5]. Seja na identificação de novas ou ainda não registradas amizades em redes sociais [6], seja na identificação de potenciais colaborações entre pesquisadores em redes de pesquisa científica [7]. A predição de links procura predizer se dois nós deveriam se relacionar baseando-se nos dados históricos e atuais referentes aos interesses e relacionamentos dos nós envolvidos[6].

Diferentemente da predição de links a predição de unlinks busca predizer quando dois nós conectados finalizarão o relacionamento entre si, ou seja, quando o link que os conecta será desfeito em um determinado momento do tempo.

Em [8, 9] os autores citam que a evolução da rede é feita tanto da adição quanto da remoção de links; portanto, para que possamos prever a evolução de redes sociais complexas é preciso predizer tanto os novos links quanto os unlinks que ocorrerão. Além da aplicação na análise da evolução destas redes, podemos utilizar a predição de unlinks com o intuito de identificar um unlink antes que ele aconteça em cenários como os que apresentamos a seguir:

- Em redes de colaboração: Consideremos que a pesquisa desenvolvida pela colaboração entre os pesquisadores x e y, que trabalham em universidade distintas, foi reconhecido por uma uma agência de fomento como um importante canal de intercâmbio de conhecimento entre as duas instituições de ensino.

O fim dessa colaboração seria prejudicial para a pesquisa em andamento e teria um impacto negativo no canal de cooperação entre as duas universidades. É importante para a agência de fomento monitorar este link e prever qualquer possibilidade de unlink para que medidas sejam tomadas a fim de preservá-lo ou promover outras formas de colaboração entre esses dois centros de pesquisa.

- Em redes de comunicação: Se for possível predizer quando um desligamento acontecerá medidas poderão ser tomadas para evitar o desligamento ou redirecionar a comunicação e manter a qualidade da transmissão.

- Em redes de comércio eletrônico: Em um um portal de e-commerce multimarcas o histórico de compras dos clientes pode ser utilizado para identificar aqueles que costumam comprar produtos de uma mesma marca. Neste caso poderíamos identificar um link entre o cliente e a marca. Em conjunto com a empresa detentora da marca, unlinks podem ser previstos e campanhas de fidelização desencadeadas para evitá-los.

A maioria das abordagens utilizadas na predição de links baseiam-se em métricas estruturais [5], como as que calculam o "menor caminho" ou o número de "vizinhos em comum" entre dois nós. Recentemente, essas métricas também foram consideradas na tarefa de predizer unlinks [10]. Os autores concluíram que apesar dos resultados obtidos terem sido melhores que os obtidos de forma aleatória, o problema da predição de unlinks se mostrou mais complicado que a predição de links, necessitando um aprofundamento da pesquisa na área.

Por outro lado, em [7] e [11] uma abordagem semântica utilizando CR$\mathcal{ALC}$[1], que considera informações sobre o domínio em questão, foi capaz de melhorar os resultados da predição de links. Apesar do levantamento da literatura realizado, não foi encontrado nenhum estudo sobre os benefícios da utilização do conhecimento do domínio na tarefa de predição de unlinks. Predizer unlinks utilizando a abordagem semântica é capaz de aprimorar os resultados?

[1] Lógica de descrição probabilísticas CR$\mathcal{ALC}$[12] (credal ALC) apresentada na seção 2.1.2

Outra característica da maioria de trabalhos relacionados à predição de links é que estes levam em consideração apenas o estado atual da rede, isto é, baseando-se em uma representação estática da rede e ignoram qualquer tipo de informação temporal que exista nela. Em [13], novas métricas estruturais que levam em conta o fator temporal foram propostas e avaliadas comparativamente com as tradicionais métricas estruturais não temporais na tarefa de predição de links e apresentaram resultados melhores.

Quando falamos de predição de unlinks e de evolução das redes, a questão temporal aparece na mudança de estados da rede de um momento do tempo para o outro. É possível representar o fator temporal na abordagem semântica? Um modelo com este tipo de informação pode melhorar a predição de unlinks?

Neste trabalho avaliaremos se o conhecimento do domínio dos objetos da rede e fatores temporais, mapeados através de uma ontologia probabilística, conseguirá predizer os unlinks que irão ocorrer na rede com mais qualidade. Para isso propomos uma abordagem semântica para a predição de unlinks que considera a informação semântica a respeito dos indivíduos do domínio e dos seus relacionamentos representados na lógica de descrição probabilística CR$\mathcal{ALC}$. Nossa abordagem foi avaliada e comparada, com sucesso, ao estado da arte em predição de unlinks utilizando uma amostra da rede de colaboração Lattes e a rede de coautoria DBLP.

Com o propósito de orientar o leitor, os capítulos dessa dissertação foram organizados da seguinte forma: a fundamentação teórica é apresentada na Capítulo 2. A proposta de solução é apresentada no Capítulo 3. No Capítulo 4 é apresentado o experimento, as bases de dados utilizadas, a metodologia empregada e os resultados obtidos, bem como a avaliação destes resultados. Alguns trabalhos relacionados são discutidos no Capítulo 5. Por fim, no Capítulo 6, são apresentadas as contribuições alcançadas, as limitações encontradas, os possíveis rumos para o futuro da pesquisa e a conclusão do trabalho.

2 Fundamentação Teórica

Neste capítulo serão apresentados alguns conceitos fundamentais para o entendimento deste trabalho. Nas seções 2.1 e 2.1.2 são conceituadas as lógicas de descrição e a lógica de descrição probabilística CR$\mathcal{ALC}$ bem como esses conceitos são utilizados para realizar a inferência de probabilidades utilizando a CR$\mathcal{ALC}$. Nas seções 2.2 e 2.3 serão apresentados conceitos relativos a redes necessários para o entendimento da proposta de predição de unlinks. Finalmente na seção 2.4 apresentamos conceitos de predição de links para que seja possível diferenciar este tipo de predição da predição de unlinks.

2.1 Lógicas de Descrição Probabilística

Lógicas descritivas (DL) formam a família das linguagens de representação que são tipicamente fragmentos de lógicas de primeira ordem [14]. O conhecimento é expresso através de indivíduos, conceitos e papéis.

Uma das lógicas de descrição mais populares é a $\mathcal{ALC}$ [15], seus construtores são: conjunção ($C \sqcap D$), disjunção ($C \sqcup D$), negação ($\neg C$), restrição existencial ($\exists r.C$), restrição universal ($\forall r.C$). Inclusões e definições de conceitos são representados por $C_1 \sqsubseteq C_2$ e $C \equiv C_2$ onde C_1 e C_2 são conceitos. O conceito $C \sqcup \neg C$ é representado por $\top$e o conceito $C \sqcap \neg C$ é representado por $\bot$.

A semântica de uma descrição é dada por um domínio $\mathcal{D}$ (um conjunto) e uma interpretação $\mathcal{I}$ (um functor). Indivíduos representam objetos através de nomes em um conjunto $N_{\mathcal{I}} = \{i_1,\ i_2,\ ...\}$. Cada conceito presente no conjunto $N_C = \{C_1,\ C_2,\ ...\}$ é interpretado como um subconjunto do domínio $\mathcal{D}$. Cada papel no conjunto $N_r = \{r_1,\ r_2,\ ...\}$ é interpretado como uma relação binária no domínio. Uma asserção assegura que um indivíduo pertence a um conceito ou que um par de indivíduos compõe um papel. Além disso:

- $(C_1 \sqcap C_2)^{\mathcal{I}} = C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$

  Ex.: $Mulher \equiv Pessoa \sqcap SexoFeminino$

O conjunto de indivíduos $Mulher$ equivale à interseção dos conjuntos de indivíduos $Pessoa$ e $SexoFeminino$.

- $(C_1 \sqcup C_2)^{\mathcal{I}} = C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$

  Ex.: $Pais \equiv Pai \sqcup Mãe$

  O conjunto de indivíduos $Pais$ equivale à união dos conjuntos de indivíduos Pai e $Mãe$.

- $(\neg C)^{\mathcal{I}} = \mathcal{D} \backslash C_1^{\mathcal{I}}$

  Ex.: $Vivo \equiv \neg Morto$

  O conjunto de indivíduos $Vivo$ equivale ao conjunto de todos os indivíduos do domínio $\mathcal{D}$ que não pertencem ao conjunto $Morto$.

- $(\exists r.C_1)^{\mathcal{I}} = \{x \in \mathcal{D},\ y \in \mathcal{D} | \exists y : (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$

  Ex.: $Avó \equiv Mãe \sqcap \exists temFilho.Pais$

  O conjunto de indivíduos $Avó$ equivale à interseção do conjunto $Mãe$ com o conjunto de indivíduos X do domínio que satisfazem a seguinte condição: existe pelo menos um y tal que $temFilho(x, y)$ é verdade e y pertence ao conjunto $Pais$.

- $(\forall r.C_1)^{\mathcal{I}} = \{x \in \mathcal{D},\ y \in \mathcal{D} | \forall y : (x, y) \in r^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$

  Ex.: $Humano \equiv Animal \sqcap \forall temParente.Humano$

  O conjunto $Humano$ equivale a interseção do conjunto $Animal$ com o conjunto de indivíduos do domínio que satisfazem a seguinte condição: para todo o y tal que $temParente(x, y)$ e y também pertence ao conjunto $Humano$.

- $C_1 \sqsubseteq C_2 = C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$

  Ex.: $Pai \sqsubseteq Pais$

  O conjunto Pai está contido no conjunto $Pais$.

- $C_1^{\mathcal{I}} \equiv C_2^{\mathcal{I}} = C_1^{\mathcal{I}} = C_2^{\mathcal{I}}$.

  Ex.: $Pai \equiv Papai$

  O conjunto de indivíduos do conjunto Pai é o mesmo do conjunto $Papai$.

2.1.1 Representação de Conhecimento em Lógicas de Descrição

Uma base de conhecimento contém as informações de um determinado domínio, ou seja, ela é a representação de um conhecimento específico. Para melhor representar este conhecimento, é necessário dividir a base em duas partes [1]:

- Conhecimento Intensional (Intensional Knowledge): conhecimento geral sobre o domínio do problema. Representa o conhecimento sobre os grupos (conjuntos) de indivíduos que apresentam as mesmas características.
- Conhecimento Extensional (Extensional Knowledge): especifica um problema particular. Representa o conhecimento sobre cada indivíduo que faz parte de um conjunto.

Em lógicas de descrição [16], o conhecimento intensional é chamado de $TBox$ e o extensional de $ABox$.

O $TBox$ contém o conhecimento intensional na forma de terminologia. Ele representa as características gerais dos conceitos, que são grupos de indivíduos semelhantes. A forma básica de declaração em um $TBox$ é a definição de conceito, que é a definição de um novo conceito em termos de outros conceitos definidos previamente.

As declarações do $TBox$ são representadas como equivalências lógicas (condições necessárias e suficientes, denotado por $\equiv$) ou como uma inclusão (condições necessárias, denotado por $\sqsubseteq$). Estas declarações possuem as seguintes características:

- Somente é permitida uma definição para cada nome de conceito.
- É recomendado que as definições sejam acíclicas no sentido de que elas não podem ser definidas em termos delas mesmas e nem de outros conceitos que indiretamente se referem a elas.

Um exemplo de um $Tbox$ é mostrado na Figura 1. Os conceitos como $Pessoa$ e $Fêmea$, no exemplo, definidos apenas pelo próprio nome, são chamados de conceitos base. O conceito $Mulher$ é declarado como o conceito resultante da interseção entre as interpretações dos conceitos $Pessoa$ e $Fêmea$, ou seja, o conceito $Mulher$

é formado pelos indivíduos que pertencem aos conceitos $Pessoa$ e $Fêmea$ simultaneamente. Este conceito é dito complexo, pois precisa de outros para ter um significado.

$$
\begin{array}{rcl}
Mulher & \equiv & Pessoa \sqcap Fêmea \\
Homem & \equiv & Pessoa \sqcap \neg Mulher \\
Mãe & \equiv & Mulher \sqcap \exists temFilho.Pessoa \\
Pai & \equiv & Homem \sqcap \exists temFilho.Pessoa \\
Pais & \equiv & Pai \sqcup Mãe \\
Avó & \equiv & Mulher \sqcap \exists temFilho.Pais \\
Avô & \equiv & Homem \sqcap \exists temFilho.Pais
\end{array}
$$

Figura 1: $TBox$ Família [1]

Para facilitar o desenvolvimento de procedimentos de raciocínio, pode-se reduzir os problemas de raciocínio com relação a um $TBox$ acíclico T para problemas com respeito ao $TBox$ vazio. Um $TBox$ vazio é um $TBox$ no qual todos os conceitos complexos são definidos apenas com a utilização de conceitos primitivos [17]. Isto permite que os algoritmos consigam encontrar mais facilmente as semelhanças entre conceitos, contradições, etc. Para se obter este $TBox$ é necessário expandir as definições de conceitos armazenados no $Tbox$ até se obter o $TBox$ vazio. Ou seja, conceitos onde cada definição que esteja na forma $A \equiv D$, D contém somente conceitos primitivos (base). Para cada conceito C, definimos a expansão de C com respeito a T como o conceito C" que é obtido de C pela substituição de cada ocorrência de um nome de símbolo A em C pelo conceito D. A Figura 2 mostra o $TBox$ vazio correspondente ao $TBox$ Família da Figura 1.

$$
\begin{array}{rcl}
Mulher & \equiv & Pessoa \sqcap Fêmea \\
Homem & \equiv & Pessoa \sqcap \neg(Pessoa \sqcap Fêmea) \\
Mãe & \equiv & (Pessoa \sqcap Fêmea) \sqcap \exists temFilho.Pessoa \\
Pai & \equiv & (Pessoa \sqcap \neg(Pessoa \sqcap Fêmea)) \sqcap \exists temFilho.Pessoa \\
Pais & \equiv & ((Pessoa \sqcap \neg(Pessoa \sqcap Fêmea)) \sqcap \exists temFilho.Pessoa) \\
 & & \sqcup((Pessoa \sqcap Fêmea) \sqcap \exists temFilho.Pessoa) \\
Avó & \equiv & (Pessoa \sqcap Fêmea) \sqcap \exists temFilho.Pais \\
Avô & \equiv & (Pessoa \sqcap \neg(Pessoa \sqcap Fêmea)) \sqcap \exists temFilho.Pais
\end{array}
$$

Figura 2: Expansão do $TBox$ Família [1]

Na Figura 2, todos os conceitos complexos foram substituídos pelos conceitos base que lhes dão origem. Neste exemplo, todos eles estão definidos em função dos conceitos Pessoa e Fêmea.

O *ABox* contém o conhecimento extensional, que especifica os indivíduos do domínio. Ele é a instanciação da estrutura de conceitos. Existem dois tipos de declarações no *ABox*:

- Declaração de Conceitos:

  $C(a)$. Declara que "a" é um indivíduo do conceito C.

  Por exemplo, $Pessoa(Ana)$.

- Declaração de Papel:

  $r(a, b)$. Declara que o indivíduo "a" está relacionado com o indivíduo "b" através do papel r.

  Por exemplo, $temFilho(Ana, João)$.

Na Figura 3, são criadas instâncias dos conceitos $Mulher$ e $Homem$. Além disso, são listados relacionamentos entre indivíduos, utilizando o papel $temFilho$, para expressar o grau de parentesco entre eles.

$Mulher(Ana)$
$Mulher(Joana)$
$Mulher(Maria)$
$Homem(Mauro)$
$Homem(Paulo)$
$Homem(Pedro)$
$temFilho(Mauro, Pedro)$
$temFilho(Mauro, Ana)$
$temFilho(Paulo, Mauro)$
$temFilho(Joana, Maria)$
$temFilho(Maria, Pedro)$
$temFilho(Maria, Ana)$

Figura 3: *ABox* da Família [1]

2.1.2 A CR$\mathcal{ALC}$

As lógicas de descrição probabilísticas são modelos com uma ótima relação entre expressividade e complexidade. Em [18] uma nova lógica probabilística chamada CR$\mathcal{ALC}$
[12] (credal $\mathcal{ALC}$) foi proposta juntamente com um algoritmo para a sua inferência, ela se baseia na lógica de descrição $\mathcal{ALC}$ e foca no desenvolvimento de métodos de inferência escaláveis.

A CR$\mathcal{ALC}$ mantém todos os construtores da $\mathcal{ALC}$ mas só permite o uso de nomes de conceitos no lado esquerdo das inclusão e definições.

Além disso, na CR$\mathcal{ALC}$ podemos ter inclusões probabilísticas $P(C_1|C_2) = \alpha$ ou $P(r) = \beta$ para os conceitos C_1e C_2 e o papel r, essas inclusões podem ser interpretadas da seguinte forma:

$$\forall x \in \mathcal{D} \ : \ P(C_1(x)|C_2(x)) = \alpha \tag{1}$$

$$\forall x \in \mathcal{D}, y \in \mathcal{D} \ : \ P(r(x,y)) = \beta \tag{2}$$

Se a interpretação de C_2 é o domínio todo então é possível representar $P(C_1|C_2) = P(C_1) = \alpha$.

Por fim, os autores adotam as seguintes suposições:

- **Aciclicidade**

  Considere as inclusões $P(C|B) = \alpha$ e $D \sqsubseteq C$. Diz se que C *utiliza* B e que D *utiliza* C; a propriedade é transitiva e portanto D "utiliza" B. Um conjunto de definições e inclusões de conceitos é chamado de terminologia; uma terminologia é dita acíclica se nenhum de seus conceitos utilizar a si mesmo. Essa propriedade é comumente imposta em lógicas de descrição [19]. Assume-se que toda terminologia em CR$\mathcal{ALC}$ é acíclica, permitindo assim que seja formado um grafo acíclico direcionado para as terminologias, possibilitando que as mesmas sejam representadas em redes Bayesianas relacionais.

- **Semântica baseada em interpretações**

  Semânticas baseadas em domínio consideram inclusões probabilísticas que são naturalmente entendidas como:

  $$P(conjunto\ de\ Cs|conjunto\ de\ Ds) = \alpha;$$

  enquanto que as baseadas em interpretação consideram inclusões probabilísticas naturalmente entendidas como:

  $$\forall x : P(C(x)|D(x)) = \alpha.$$

  Esta última semântica é utilizada em CR$\mathcal{ALC}$ pois permite fazer inferências sobre $P(A(i)|B(j))$ para os conceitos A e B respectivamente instanciados para os indivíduos i e j enquanto que uma semântica baseada em domínio atribuiria o valor 0 ou 1 para a interpretação, ficando presa no problema de inferência direta. Note que não há contradição entre $\forall x : P(C(x)) = \alpha$ e uma observação $C(a) = verdadeiro$ já que $P(C(a)|C(a)) = 1$ enquanto ainda temos $P(C(a)) = \alpha$. Assume-se que os indivíduos são designadores rígidos em relação às interpretações, ou seja, um indivíduo corresponde ao mesmo elemento do domínio para todas as interpretações [20], como será visto nos parágrafos a seguir.

- **Independência e Condição de Markov**

  Lógicas de descrição probabilísticas têm uma propriedade de independência definida por uma condição de Markov que restringe os valores de probabilidade e decompõe os modelos em pequenas partes isoladas. Nessas lógicas não há uma sintaxe que expresse independência; independência é extraída diretamente da estrutura das fórmulas, fato comum a muitas lógicas [21, 22]. A lógica CR$\mathcal{ALC}$ também assume essa posição, sendo que a independência é representada pela seguinte condição de Markov:

Para todo conceito C em uma terminologia e para cada x no domínio $\mathcal{D}$, $C(x)$ é independente de todas as afirmações que não utilizem $C(x)$, dadas as afirmações que utilizem diretamente C.

Para todo papel r e para todo (x, y) em $\mathcal{D}\times\mathcal{D}$, $r(x, y)$ é independente de todas as afirmações que não utilizem $r(x, y)$, dadas as afirmações que utilizem diretamente r.

- **Homogeneidade**

  Uma terminologia pode não especificar uma única distribuição de probabilidades sobre as interpretações. Por exemplo: $P(C|B \sqcup \exists r.D) = \alpha_1$ não garante que $P(C|B \sqcap \exists r.D)$ seja restrito a um único valor. Por isso, assume-se a seguinte condição de homogeneidade:

  Considere um conceito C com pais $C_1, ..., C_m$. Para cada conjunção de m conceitos $\pm C_i$, tem-se $P(C| \pm C_1 \sqcap \pm C_2 \sqcap \cdots \sqcap \pm C_m) = \gamma$. [2]

  Com essa condição, cada terminologia em CR$\mathcal{ALC}$ pode ser representada como um grafo acíclico direcionado onde cada nó está associado a uma relação. Se cada probabilidade for especificada por um valor preciso, então a terminologia pode ser representada por uma rede Bayesiana relacional, sendo que a proposicionalização da mesma gera uma rede Bayesiana. Para casos onde nem todas as probabilidades são precisamente especificadas, a proposicionalização gera uma chamada rede credal [23].

- **Unicidade**

  Adota-se a suposição de nomes únicos: nomes distintos de indivíduos correspondem a elementos distintos do domínio. Adota-se também a seguinte condição de unicidade:

[2] *O sinal "$\pm$" indica que o conceito pode aparecer negado ou não.*

– Cada conceito C pertence a uma de duas categorias: ou C é especificado por uma definição, ou C tem um pai D e são feitas inclusões a respeito de D e $\neg D$. Para o ultimo caso há três opções:

  i $C \sqsubseteq D$ e $P(C|D) = \alpha$;

  ii $C \sqsubseteq \neg D$ e $P(C|\neg D) = \alpha$;

  iii $P(C|D) = \alpha'$ e $P(C|\neg D) = \alpha''$.

– Para cada papel r é feita uma atribuição de probabilidade $P(r) = \alpha$, cuja semântica é $\forall x, y : P(r(x, y)) = \alpha$.

2.1.3 Inferência utilizando CR$\mathcal{ALC}$

Em consequência da condição de Markov, a CR$\mathcal{ALC}$ pode ser usada para especificar redes bayesianas relacionais de forma simples com elementos de lógica.

Como é assumido que toda a terminologia é acíclica, isto é, nenhum conceito utiliza ele mesmo, onde "usar" é encerramento transitivo de "usar diretamente", podemos dizer que C_1 usa diretamente C_2 se C_2 aparece no lado direito de uma inclusão/definição, ou no lado condicionante de uma inclusão probabilística esta suposição permite que representemos qualquer terminologia $\mathcal{T}$ através de um grafo acíclico direcionado. Este grafo, denotado por $\mathcal{G}(\mathcal{T})$, possui cada nome de conceito e cada nome de papel representado como nó do grafo, e se o conceito C_1 usa diretamente o conceito C_2, isto é, se C_1 aparece no lado esquerdo e C_2 no lado direito de uma inclusão/definição então C_2 é pai de C_1 em $\mathcal{G}(\mathcal{T})$. Cada restrição existencial $\exists r.C$ e cada restrição de valor $\forall r.C$ é adicionada ao grafo $\mathcal{G}(\mathcal{T})$ como um nó com uma aresta de r e C para cada nó de restrição que os use diretamente. Cada nó de restrição é um nó determinístico pois seu valor é totalmente determinado pelos seus pais.

Consideremos a seguinte terminologia como um exemplo:

$$Researcher \equiv Person \sqcap \exists hasPublication.BibItem$$

$$P(Person) = 0.2$$

$$P(BibItem) = 0.6$$

$$P(hasPublication) = 0.1$$

Seu grafo seria representado pela figura 4.

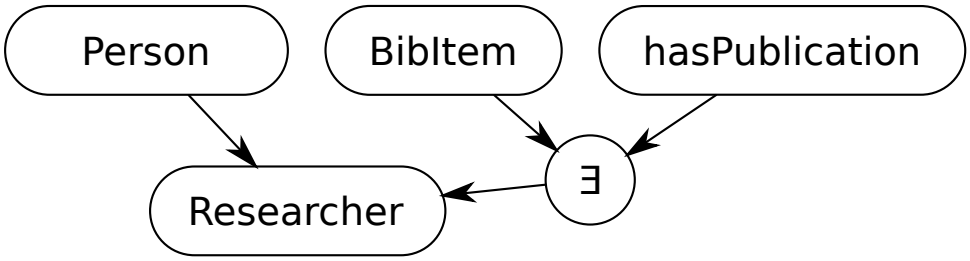


Figura 4: Grafo $\mathcal{G}(\mathcal{T})$

A semântica da CR$\mathcal{ALC}$ é baseada em medidas de probabilidade sobre o espaço de interpretações, para um domínio fixo. Para garantir que a terminologia especifica uma medida de probabilidade única, algumas suposições são adotadas:

- o domínio é finito, fixo e conhecido
- indivíduos possuem nome únicos e a rigidez [24]
- apenas um nome de conceito pode aparecer à esquerda de qualquer inclusão/definição e no lado condicional de qualquer inclusão probabilística
- a condição de Markov impõem a independência da proposicionalização de qualquer conceito/papel condicional na proposicionalização de seus correspondentes pais no grafo $\mathcal{G}(\mathcal{T})$ [12].

Levando em consideração estas suposições, um conjunto de sentenças $\mathcal{T}$ em CR$\mathcal{ALC}$ define uma rede bayesiana relacional [25] cujo grafo subjacente é exatamente $\mathcal{G}(\mathcal{T})$, como ilustrado na Figura 5.

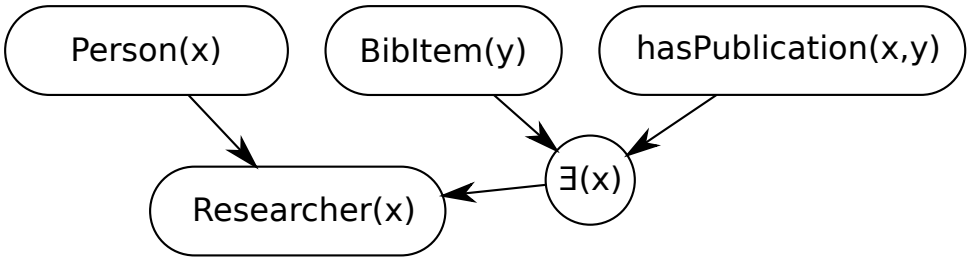


Figura 5: Rede Bayesiana definida a partir de $\mathcal{T}$

Considerando a terminologia $\mathcal{T}$, o domínio $\mathcal{D} = \{bob,\ paper,\ x\}$, e o conjunto de asserções ($ABox$) $\mathcal{A} = \{Person(bob)$, $Resear\text{-}cher(bob)$, $BibItem(paper)$, $hasPublication(bob, paper)\}$, as suposições discutidas induzem uma medida de probabilidade única sobre o conjunto de todas as asserções pois induzem uma rede bayesiana sobre as variáveis indicadoras das asserções que é apresentada na figura 6. Os nomes foram abreviados para economizar espaço: b, p e x representam os indivíduos bob, paper e x respectivamente enquanto que hP, P, BI e R representam o papel e os conceitos $hasPublication$, $Person$, $BibItem$ e $Researcher$ respectivamente.

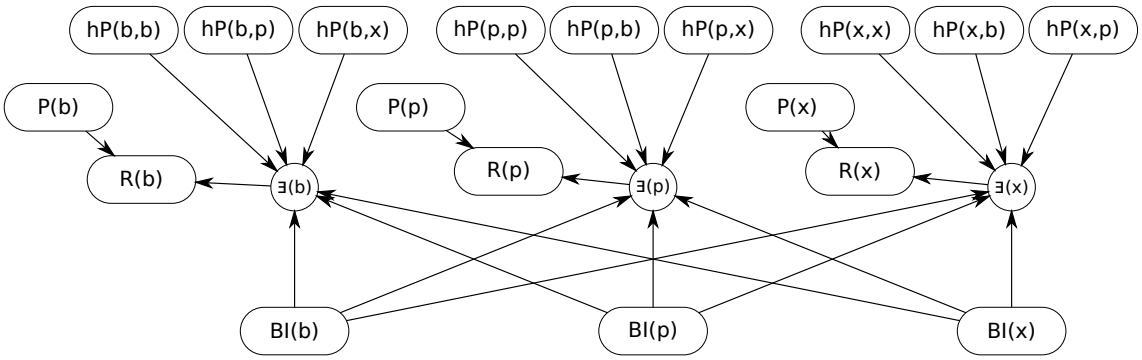


Figura 6: Rede Bayesiana sobre as variáveis indicadoras das asserções produzida pela proposicionalização da terminologia $\mathcal{T}$. Três slices foram criados um para cada indivíduo.

Inferências tais como $P(Researcher(bob)|\mathcal{A})$ para o $ABox$ $\mathcal{A}$, podem ser computadas proposicionalizando a terminologia ($TBox$) a fim de gerar uma rede bayesiana

onde um slice é gerado para cada indivíduo. Para domínios com uma grande quantidade de indivíduos realizar a inferência probabilística exata geralmente é muito difícil mas algoritmos variacionais capazes de realizar inferências aproximadas estão disponíveis na literatura [12].

2.2 Redes Complexas

As redes complexas modelam sistemas dinâmicos e possuem propriedades e características que têm sido estudadas pela academia. O conjunto de áreas do conhecimento que se utilizam da Teoria de Redes Complexas é grande e heterogêneo. Há pesquisas envolvendo redes complexas em um leque de campos que vai de Artes a Zoologia, passando por Linguística e Psicologia[26]. Além de ser particularmente interessante como ciência pura, o estudo sobre Redes Complexas é importante pelo impacto de sua aplicação no mundo real como na predição de links.

A pesquisa em redes complexas é multidisciplinar e engloba conceitos de teoria dos grafos, estatística e sistemas complexos para apoiar a caracterização, análise e modelagem dos mais variados fenômenos [27]. Uma rede pode ser definida como um conjunto de itens conectados por relações existentes entre si, nos quais os itens são os vértices (nós) e suas conexões são os links (arestas). Formalmente, uma rede $G = (V, E)$ contém um conjunto de vértices V e um conjunto de arestas E.

Com relação à forma como os nós e arestas armazenam informações importantes ao entendimento da rede como um todo, podemos citar as seguintes características comumente encontradas:

- multimodal: quando os nós possuem atributos, dessa forma podemos ter numa rede tipos diferentes de nós.
- multirrelacional (multi-plex): quando as informações estão presentes nas ligações, o que permite a existência de diferentes tipos de links [28];
  - as arestas são ponderada ou não: as arestas possuem um peso ou valor associado a elas. Numa rede de leitores de livros uma aresta que liga um

leitor a um livro pode ter associada a ela um valor referente à nota da avaliação do livro lido pelo leitor .

  - as arestas possuem uma data: representa o momento que a relacionamento ocorreu ou se iniciou. Em um rede formada pela troca de mensagens entre usuários a aresta possui um atributo data que sinaliza o momento em que a mensagem foi enviada.
  - as arestas possuem outros atributos: levando em consideração o exemplo da rede de troca de mensagens, a própria mensagem poderia ser um atributo da aresta.

- se suas arestas são direcionadas ou não: quando uma aresta é direcionada então $r(x, y) \neq r(y, x)$. Em uma rede social para microblogging como o Twitter o relacionamento "*x segue y*" não implica que "*y segue x*". Já em uma rede científica de colaboração um relacionamento "*x colabora com y*" é bidirecional quando x e y são coautores de um mesmo artigo.

- se podem existir múltiplas arestas entre dois nós: se estas arestas não possuírem nenhum atributo então a quantidade de arestas pode ser interpretada como o peso deste relacionamento ou se cada um destas arestas podem possuir uma data diferente podemos então interpretar como um sequência de eventos e avaliar a duração deste relacionamento.

Quanto ao processo de formação das redes, podemos destacar os modelos a seguir:

- Rede Aleatória [29]: É o modelo de formação de redes mais simples que uma rede complexa pode assumir. A formação da rede segue um processo aleatório, onde partindo de um conjunto de N vértices, são adicionadas arestas que ligam os vértices uns aos outros com probabilidade p. Esse modelo gera grafos aleatórios com N vértices e k arestas, denominados grafo aleatório ER, definido como $G_{N,k}^{ER}$. Inicialmente com N vértices desconectados, o modelo ER é obtido conectando-se os vértices selecionados aleatoriamente até o número de arestas do grafo ser igual a k. Acredita-se que o processo de construção da rede seja

aleatório no sentido de que vértices se agregam aleatoriamente. Com base nessa premissa, concluíram que todos os vértices de uma determinada rede têm aproximadamente a mesma quantidade de conexões e as mesmas chances de receberem novas ligações. Segundo [30], quanto mais complexa for a rede, maiores serão as chances dela ser aleatória. Outra implicação deste modelo é que para uma rede grande, ainda que os links estejam dispostos de maneira aleatória, boa parte dos vértices terão aproximadamente o mesmo grau, o que significa que a longo prazo, nenhum vértice será favorecido ou isolado [31, 27].

- Rede Mundo Pequeno [32]:

  Em 1967, [33] realizou um experimento, se uma carta fosse entregue a um indivíduo, que não fosse o destinatário, e se ele ão conhecesse o destinatário que ele a repassasse a uma pessoa qualquer de suas relações que tivesse maior chance de conhecer e assim por diante, em aproximadamente seis passagens esta carta chegaria ao destinatário. Esse resultado é uma demonstração direta do efeito pequeno-mundo, em que o caminho percorrido pela carta, partindo de um indivíduo qualquer até o destinatário, é mínimo. Com esse experimento o conceito dos seis graus de separação entre pessoas.

  Com base no conceito de Seis Graus de Separação, [34] propôs que, dentro de uma rede de relacionamentos sociais, a sociedade seria composta por grupos altamente conectados, existindo apenas poucos vínculos externos (ligações fracas) que conectam esses grupos a outros, evitando o isolamento deles com o resto da rede.

  O modelo de formação proposto por [32], conhecido como modelo mundo pequeno, é uma alternativa ao modelo randômico de [29] e consiste inicialmente em criar grupos densos (grupos de vértices altamente conectados) na rede, para depois inserir novos links ligando vértices aleatoriamente escolhidos. Esses links oferecem atalhos cruciais entre vértices distantes, evitando o isolamento dos grupos e diminuindo o grau de separação entre os vértices da rede.

- Rede de Livre Escala [30]:

  Ao estudar as propriedades da Web como uma rede, descobriram uma característica não observada pelos modelos anteriormente apresentados: a existência de alguns poucos vértices altamente conectados, os quais chamaram de hubs, e muito vértices com poucas conexões. Os hubs estão presentes em diversas redes complexas e exercem um papel importante na disseminação de informações dentro da rede.

  Esse modelo foi chamado de livre de escala ou sem escala e são aquelas em que a distribuição de links na rede deveria seguir uma lei de potência, conforme a Equação 2.2, onde γ e o expoente do grau, que para a maioria dos sistemas varia em torno de 2 e 3.

$$p_k \sim k^{-\gamma} \tag{3}$$

  Em [32] os autores demonstraram que grande parte das redes reais apresenta como uma característica bem específica, a conexão preferencial, que é a tendência de um novo vértice se conectar a um vértice da rede que tem um grau elevado de conexões, também conhecida como a ideia de que o rico fica mais rico. Assim, o que acontece no processo de crescimento das redes é que não são os vértices mais antigos os mais propensos a adquirirem links ao longo do tempo, senão, são os novos membros os mais favoráveis a se conectarem com aqueles que apresentam mais conexões na rede. Logo, os vértices mais bem conectados serão escolhidos com mais frequência e crescerão mais rapidamente que outros vértices menos conectados.

  Conforme apresentado em [31], essas redes têm sido observadas em vários sistemas, por exemplo, na internet, na Web, em redes de metabolismos e em redes de citações de artigos científicos.

Tomemos como exemplo a coleção de Redes Koblenz (Konect[3]), um projeto do Instituto de Ciências e Tecnologias Web da Universidade de Koblenz-Landau com o objetivo de coletar grandes conjuntos de dados que representem redes de todos os tipos para a realização de pesquisas na área de mineração de rede. Atualmente, as fontes de dados já coletadas somam 189 amostras de redes, com características diversas, extraídas de grandes redes sociais como: Twitter, Flickr, Digg, Amazon, Netflix, Wikipedia e outras. De acordo com as características apresentadas, as redes que compõem a coleção foram classificadas da seguinte forma:

- 41 possuem arestas não direcionadas
- 69 possuem arestas direcionadas
- 79 são bipartites
- 80 possuem arestas que não são ponderadas
- 80 permitem múltiplas arestas entre dois nós
- 71 possuem arestas com data

2.3 Redes Evolutivas

Uma rede evolutiva é uma rede que evolui, ou se modifica, conforme o tempo passa. Consideremos que $G(V, E)$ é um grafo representando uma rede social evolutiva onde V é o conjunto de nós e E o conjunto de arestas. Cada aresta em E é representada por uma tripla (u, v, t) indicando que os nós u e v, $\{u, v\} \subset V$, possuem um relacionamento entre si e que este relacionamento é presente no tempo t [13]. Dessa forma, poderíamos dizer que o grafo G pode ser dividido em diversos subgrafos organizados de acordo com o intervalo de tempo considerado: hora, dia, mês, ano, etc. Na Figura 7 é apresentado um exemplo de rede evolutiva, onde o grafo G foi dividido em subgrafos, um para cada intervalo $i \in I_E$ que é o conjunto de intervalos obtidos ao dividir as arestas em E.

[3]http://konect.uni-koblenz.de/

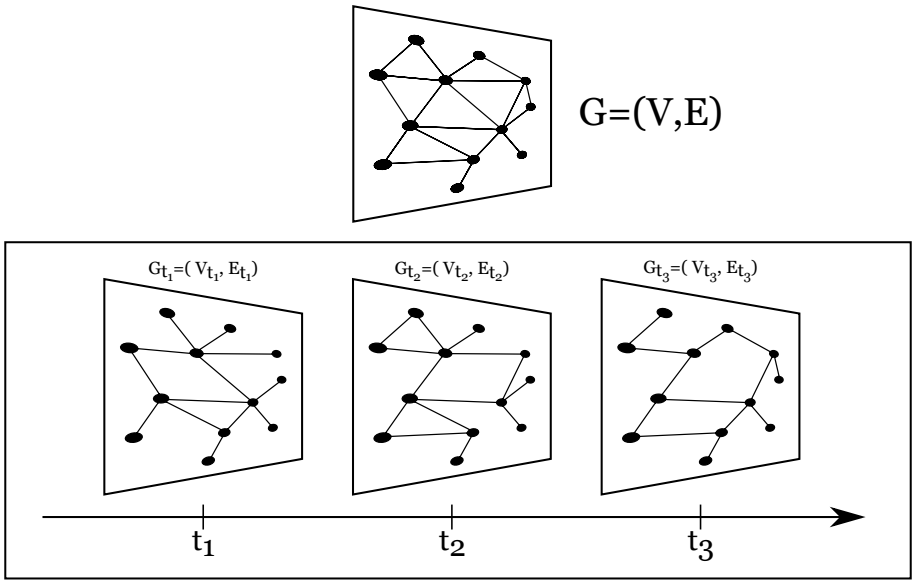


Figura 7: Exemplo de uma Rede Evolutiva $G = G_{t_1} \cup G_{t_2} \cup G_{t_3}$

$$G = \bigcup_{i \in I_E} G_{t_i} \qquad onde\ G_{t_i} = (V_{t_i}, E_{t_i}) \tag{4}$$

Podemos listar algumas características deste tipo de rede:

- o fator temporal permite que a rede tenha um histórico dos relacionamentos e das iterações entre os seus nós, sendo possível determinar o início, fim e a duração desses relacionamentos.

- em determinados domínios poderemos ter triplas (u, v, t) com semânticas diferentes representando diferentes tipos de relacionamentos entre os nós como por exemplo: uma tripla (u_1, v_1, t_1) que representa u_1 *orienta* v_1 em t_1 e outra tripla (u_2, v_2, t_2) que representa u_2 *coautor* v_2 em t_2 sendo que $u_1 = u_2$, $v_1 = v_2$ e $t_1 = t_2$.

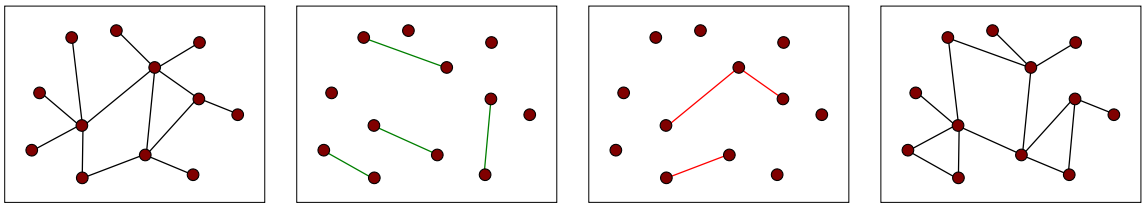
- dependendo da unidade de tempo utilizada poderemos ter mais de uma tripla (u, v, t), com a mesma semântica e no mesmo intervalo, indicando múltiplos relacionamentos entre u e v que podem ser representados no grafo através de várias arestas entre u e v ou através de uma única aresta que possua um atributo peso indicando a quantidade de relacionamentos com a mesma semântica que ocorreram no intervalo t.

- alguns relacionamentos podem ser atemporais ou então terem um início mas não um fim.

Dado um conjunto de triplas em E nos tempos t_1 e t_2 podemos dizer que E_{t_1} evolui para E_{t_2}através da adição de novas triplas $E^+_{t_{(1,2]}}$e da remoção de triplas $E^-_{t_{(1,2]}}$ dentro deste intervalo. Dessa forma podemos definir essa evolução através da equação 5:

$$E_{t_2} = (E_{t_1} - E^-_{t_{(1,2]}}) \cup E^+_{t_{(1,2]}} \tag{5}$$

A Figura 8 ilustra como uma rede evolui do tempo t_1 para o tempo t_2: em (a) temos o conjunto de arestas E_{t_1} no tempo t_1, (b) o conjunto de arestas $E^+_{t_{(1,2]}}$ é adicionado e (c) o conjunto $E^-_{t_{(1,2]}}$ é removido para que tenhamos como resultado (d) o conjunto de arestas E_{t_2}no tempo t_2.



(a) E_{t_1} (b) $E^+_{t_{(1,2]}}$ (c) $E^-_{t_{(1,2]}}$ (d) E_{t_2}

Figura 8: Evolução de uma rede do intervalo t_1 para o intervalo t_2
(figura baseada na figura 2 de [2])

2.4 Predição de Links

A tarefa de predizer links pode ser definida da seguinte forma [6]: dada uma rede $G = (V, E)$ representada como um grafo e constituída por um conjunto de nós e um conjunto de arestas, onde uma aresta representa uma interação entre os nós da rede, devemos predizer as arestas que irão surgir de acordo com as arestas que já existem na rede. Precisamos predizer se um link entre dois nós desconectados, a e b, deve existir dados os nós e arestas que já existem na rede.

Vários tipos diferentes de ferramentas são utilizadas para realizar a predição de links, algumas como a fatoração de matrizes utilizam-se do grande volume de dados disponíveis para isso, outras estão diretamente relacionadas à existência de links entre os nós. É possível utilizar classificadores que, baseados em atributos da rede e outras medidas, são capazes de classificar os possíveis links em verdadeiros ou falsos [5], outros podem utilizar-se de classificação coletiva de acordo com todo o conjunto de possíveis links [35]. Muitas técnicas de predição de links são baseadas em medir-se a proximidade ou similaridade dos nós da rede [6, 36].

Outras abordagens consideram atributos semânticos para calcular a similaridade semântica entre dois nós e podem ser bastante úteis na predição dos links desprezados por medidas topológicas simples [37]. Uma forma de capturar esta semelhança semântica é através de documentos diretamente relacionados aos nós em questão. Um simples exemplo de similaridade semântica que pode ser utilizada é a contagem de palavras em comum entre dois autores [38]. Um método mais sofisticado faz uso de técnicas bem conhecidas como a representação vetorial de atributos TFIDF e a medida do coseno para calcular a similaridade [37].

Abordagens para a predição de links podem ser entendidas não apenas levando em consideração as ferramentas utilizadas, mas também pela análise do modelo que é utilizado para representar a rede como um todo. Tipicamente assume-se algum tipo de mecanismo probabilístico que explique, minimamente, a existência das arestas em conjunto com o conhecimento específico do domínio, como teorias a respeito das relações entre pessoas [39, 31].

Outra forma de probabilisticamente predizer mudanças na estrutura de uma rede é através de modelos baseados no grafo tais como campos aleatórios de Markov ou redes bayesianas [40]. Apesar de estas linguagens serem bem adaptadas para expressar as relações independentes entre um conjunto fixo de variáveis randômicas, quando nós e links devem ser tratados dentro de grafos é melhor considerar linguagens que possam especificar campos aleatórios de Markov e redes bayesianas de acordo com estruturas relacionais. De fato muitas propostas para a predição de links recorrem a

essas linguagens como pode ser conferido nos trabalhos de [35] e [41]. A presença da estrutura relacional permite que propriedades individuais de nós, links e comunidades sejam representadas e a partir destes dados calcular e estimar a probabilidade de um link específico .

Em [7], esta estratégia de modelagem foi seguida utilizando-se da lógica de descrição probabilística CR$\mathcal{ALC}$. O interesse em modelos baseados em lógicas de descrição é justificado, dados os resultados recentes da importância das ontologias na organização da informação que pode ser utilizada para a predição de links [42, 43, 44]. Enquanto outras implementações de predição de links usualmente focam em um tipo de atributo, a utilização da CR$\mathcal{ALC}$ mostrou-se capaz de misturar diferentes atributos tais como semânticos, numéricos e topológicos. Ser uma solução versátil não faz com que a sua modelagem seja mais fácil que outras soluções, mas por ser uma abordagem nova ainda há espaço para a sua evolução e a necessidade de mais experimentação.

2.4.1 Métricas de similaridade

As principais métricas utilizadas para calcular a similaridade entre dois nós e que são aplicadas na predição de links são apresentadas a seguir:

- Vizinhos em comum (Common Neighbors) [45]

  É a quantidade de vizinhos (Γ) em comum entre dois nós (r_1 e r_2).

$$score(r_1, r_2) = \mid \Gamma(r_1) \cap \Gamma(r_2) \mid \tag{6}$$

- Coeficiente de Jaccard (Jaccard coefficient) [46]

  É uma normalização da medida "Vizinhos em comum", comumente utilizada na recuperação de informação [46], para medir a probabilidade de que r_1 e r_1 tenham uma característica f em comum. Neste artigo, consideramos que f se refere a quantidade de vizinhos (Γ) em comum.

$$score(r_1, r_2) = \frac{\mid \Gamma(r_1) \cap \Gamma(r_2) \mid}{\mid \Gamma(r_1) \cup \Gamma(r_2) \mid} \tag{7}$$

- Adamic/Adar [47]

  Refina a ideia dos vizinhos em comum considerando os vizinhos menos populares mais relevantes no somatório.

$$score(r_1, r_2) = \sum_{z \in V} \frac{1}{log(\Gamma(z))} \quad , \; onde \; V = \Gamma(r_1) \cap \Gamma(r_2) \tag{8}$$

- Comprimento do menor caminho (shortest path distance) [45]

  Segue a ideia de que os amigos dos amigos são uma boa sugestão para novos links. A similaridade é calculada encontrando-se o comprimento do menor caminho na rede que liga os nós r_1 e r_2.

- Conexão preferencial (preferential attachment) [48]

  Propõem que a probabilidade de um nó se conectar a outro baseia-se na quantidade de vizinhos que eles possuem. Quanto maior a quantidade de vizinhos (Γ), maior a probabilidade de conexão. Pode ser calculada pela multiplicação das quantidades, mas também existem artigos onde a medida é calculada pela soma das quantidades [38].

$$score(r_1, r_2) = \Gamma(r_1) \cdot \Gamma(r_2) \quad ou \quad score(r_1, r_2) = \Gamma(r_1) + \Gamma(r_2) \tag{9}$$

- Coeficiente de agrupamento (clustering coefficient) [45]

  É calculado pela multiplicação ou soma dos coeficientes de agrupamento dos dois nós em questão. Evidências [Watts and Strogatz 1998] sugerem que a probabilidade de surgir um novo link entre dois nós que pertencem a um mesmo grupo da rede social é maior do que entre dois nós escolhidos aleatoriamente. As fórmulas, da métrica (equação 4), bem como do coeficiente de agrupamento

de um determinado nó (equação (5)), são:

$$score(r_1, r_2) = cc(r_1) \cdot cc(r_2) \quad ou \quad score(r_1, r_2) = cc(r_1) + cc(r_2) \tag{10}$$

$$cc(v) = \frac{3 \cdot qtd.\,de\,triangulos\,adjacentes\,à\,v}{qtd.\,de\,possíveis\,triangulos\,adjacentes\,à\,v} \tag{11}$$

- Katz [49]

  Soma a quantidade de todos os caminhos entre r1 e r2 de acordo com o comprimento do caminho, exponencialmente amortecido por este comprimento. Dessa forma os caminhos mais curtos são considerados mais relevantes que os compridos. De acordo com [47] caminhos com comprimento igual a três ou maiores pouco contribuem no somatório. A seguir apresentamos sua fórmula onde l é o comprimento dos caminhos, que deve variar de 1 até k, sendo k normalmente igual a 2, β é o fator de amortecimento necessário para que os caminhos com comprimento menor sejam considerados mais relevantes no cálculo e $paths(r_1, r_2, l)$ é a função que retorna a quantidade de caminhos encontrada entre r_1 e r_2 com o comprimento igual a l.

$$score(r_1, r_2) = \sum_{l=1}^{k} \beta^l \cdot paths(r_1, r_2, l) \tag{12}$$

2.4.2 Métodos de classificação temporais

- *Baseada* em eventos temporais (Event Based) [13]

  Um evento temporal é uma ação que altera ou mantém o estado de um par de nós de desconectados para conectados. Dessa forma podemos classificar os eventos temporais em três tipos:

  - Conservativo

    Um evento conservativo ocorre quando um relacionamento, entre dois nós, que existia em $t-1$ volta a ocorrer em t.

$$C(u,v,t) = \begin{cases} c & se\,(u,v) \in E_{t-1} \cap E_t \\ 0 & senão. \end{cases} \tag{13}$$

– Inovativo

Um evento inovativo ocorre quando um relacionamento que não existia em $t-1$ ocorre em t.

$$I(u,v,t) = \begin{cases} i & se\,(u,v) \in E_t - E_{t-1} \\ 0 & senão. \end{cases} \tag{14}$$

– Regressivo

Um evento regressivo ocorre quando um relacionamento, entre dois nós, que existia em $t-1$ deixa de existir em t.

$$R(u,v,t) = \begin{cases} r & se\,(u,v) \in E_{t-1} - E_t \\ 0 & senão. \end{cases} \tag{15}$$

- Crescimento/Declínio da rede (Growth/Decay Based) [2]

  Considera que um nó que ao longo de vários intervalos consecutivos do tempo adiciona novas ligações sem remover as antigas, possui um padrão de *crescimento* da sua rede pessoal. Enquanto que um nó que não adiciona novas ligações e vem gradativamente perdendo as antigas conexões constitui uma rede em *declínio*. Nós conectados que apresentam redes em declínio estão mais propensos a cortarem ligações entre si do que os nós que estão expandindo suas redes.

- *Baseada na Estabilidade* (Stability Based) [2]

  Um nó que mantém as mesmas conexões através do tempo, sem adicionar novas ligações ou remover as antigas, pode ser classificado como um nó *estável*. Inversamente, um nó *instável*, apesar de manter o tamanho da sua rede pessoal, adiciona e remove conexões fazendo com que a sua rede pessoal mude a cada

intervalo de tempo analisado. Nós instáveis são mais propensos a cortarem ligações com seus nós vizinhos do que os nós estáveis.

3 Proposta de Solução

A seguir apresentamos a proposta de solução desta dissertação, começando pela definição do que é a predição de unlinks para em seguida apresentar a metodologia empregada para realizar esta predição.

3.1 Predição de Unlinks

A predição do unlink é a tentativa de predizer a mudança do estado "conectado" de um link, entre dois nós na rede, para o estado "desconectado", representada na Figura 9 pelo item 1. Se a predição é positiva o link entre os dois nós é desfeito e passa ao estado "desconectado". Se a predição é negativa o link entre os dois nós se mantém e permanece no estado "conectado". É importante ressaltar que a predição do unlink não é o mesmo que a predição negativa de um link, assinalada na Figura 9 pelo item 2.

A Figura 9 apresenta as possíveis mudanças de estado dos links que as tarefas de predição de links e de predição de unlinks buscam predizer.

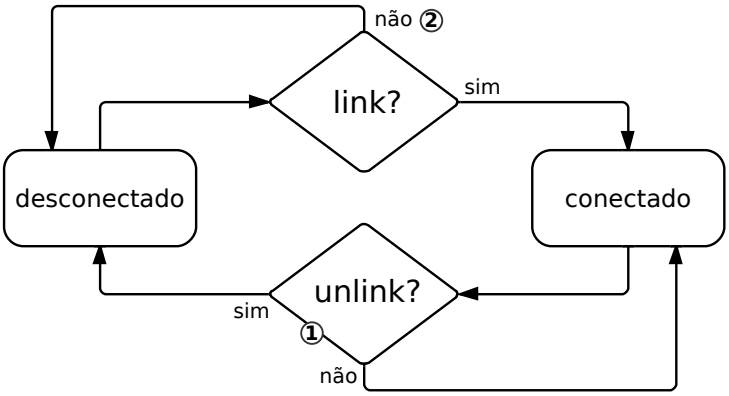


Figura 9: Mudança de estados entre relacionamentos

Por se tratar de um problema diferente, as bem sucedidas abordagens estruturais utilizadas para a predição de links precisam ser adequadamente avaliadas para a predição de unlinks. Entretanto, durante o desenvolvimento deste trabalho, e o entendimento do racional de algumas das principais métricas estruturais, foram percebidos alguns possíveis problemas que podem desfavorece-las.

Um destes problemas se deve ao fato de que links podem se desfazer abruptamente por diversas razões que métricas estruturais não são capazes de predizer. Por exemplo, uma relação aluno-orientador, embora forte durante todo o tempo que dura a orientação, tem alguma probabilidade de cessar após o aluno completar sua formação. No entanto, preditores baseados em abordagens estruturais ou topológicas provavelmente garantirão a continuidade do vínculo com base nas conexões em comum com outros pesquisadores e no trabalho em conjunto realizado recente. Em resumo, o unlink pode ocorrer mesmo entre nós considerados próximos.

Outros problemas inerentes das métricas estruturais são o desconhecimento da semântica dos nós e o desconhecimento da semântica e da direção dos relacionamentos. Esta falta de semântica não influencia a predição de links em redes single-plex[36], onde existe apenas um tipo de relacionamento, e em redes onde a direção é ignorada, mas em redes multi-plex, onde podemos ter vários tipos de relacionamentos com semânticas diferentes entre si e em redes onde a direção do link é importante, pode haver problemas como:

Dependendo do domínio em questão e do tipo de link que queremos predizer, alguns tipos de links existentes na rede e que são levados em consideração nos cálculos das métricas podem ter pouco ou nenhum valor na predição. Em uma rede social como o Facebook onde podemos encontrar links como “x gosta de y” ou [50][51].

Portanto, acreditamos que o conhecimento do domínio das redes seja necessário para melhor predizer unlinks. Neste trabalho, propomos uma abordagem através da modelagem do domínio de redes de coautoria utilizando ontologias probabilísticas representadas na linguagem de descrição probabilística CR$\mathcal{ALC}$ e através dos

modelos criados inferimos as probabilidades do unlink ocorrer.

Como formalizado nas seções 2.1 e 2.3, desejamos construir uma rede onde os nós são indivíduos de um dos conceitos definidos em uma ontologia O e as arestas são instâncias de um dos papéis definidos em O. A rede G pode então ser construída a partir das asserções dos conceitos e papéis sobre os indivíduos do domínio. Por exemplo, em uma rede de coautoria, asserções do conceito $Person$ são representadas como nós e as asserções do papel $sharePublication$ são representadas por arestas.

Uma ontologia probabilística O pode modelar o domínio com outros conceitos e papéis além daqueles cujas asserções se fazem presentes no $ABox$. Dessa forma, apesar de asserções referentes ao conceito $Researcher$ não existirem no $ABox$, é possível deduzi-las a partir da sua definição na terminologia $TBox$ e das asserções dos conceitos relacionados a $Researcher$.

$$
\begin{array}{lll}
TBox: & P(Person) & = 0.3 \\
 & P(BibItem) & = 0.3 \\
 & P(sharePublication) & = 0.22 \\
 & Researcher & \equiv Person \sqcap \exists hasPublication.BibItem \\
 & P(Coauthor | Researcher \sqcap \exists sharePublication.Researcher) & = 0.91
\end{array}
$$

$$
\begin{array}{ll}
ABox: & Person(john), Person(ann), \\
 & BibItem(p1), BibItem(p2), BibItem(p3), \\
 & hasPublication(john, p1), hasPublication(john, p2), \\
 & hasPublication(ann, p3), hasPublication(ann, p2), \\
 & sharePublication(john, ann)
\end{array}
$$

Figura 10: Um exemplo de ontologia probabilística para o domínio de coautoria.

Como a rede G é uma rede evolutiva, existem asserções a respeito dos indivíduos do domínio em diferentes intervalos do tempo. Para representar esta realidade assumimos que existem vários ABoxes, cada um referente a um período de tempo distinto dos outros. Na Figura 11 apresentamos um exemplo de ontologia probabilística, representando um domínio acadêmico, com três ABoxes representando os momentos no tempo t_1, t_2 e t_3.

$$
\begin{array}{lll}
TBox: & P(Student(u)) & = 0.70 \\
 & P(Professor(u)) & = 0.30 \\
 & P(isAdvisorOf(u,v)|Professor(u) \sqcap Student(u)) & = 0.22 \\
 & P(advisoryEnds(u,v)|isAdvisorOf(u,v)) & = 0.10 \\
 & Advisor \equiv Professor \sqcap \exists isAdvisorOf.Student & \\
 & P(unlink(u,v)|advisoryEnds(u,v)) & = 0.67
\end{array}
$$

$ABox_{t_1}$	$ABox_{t_2}$	$ABox_{t_3}$
$Professor(john)$	$Professor(john)$	$Professor(john)$
$Student(ann)$	$Student(ann)$	$Student(ann)$
$isAdvisorOf(john, ann)$	$isAdvisorOf(john, ann)$	
	$advisoryEnds(john, ann)$	

Figura 11: Uma ontologia com três ABoxes um para cada período de tempo.

Para predizer o unlink em um determinado momento t_2 as asserções existentes no o $ABox_{t_2}$e nos $ABoxes$ anteriores a t_2 são utilizadas para complementar o $ABox_{t_2}$.

Por ser um exemplo simples, com poucos conceitos e papéis, a predição do unlink baseada apenas no fato de que a orientação chegou ao fim pode parecer uma solução óbvia para um problema trivial, mas se considerarmos que uma ontologia mais complexa será formada por muitos outros conceitos e papéis e que poderemos mapear outros motivos para que o unlink ocorra, como a mudança do aluno para outro estado ou a transferência do professor para outra instituição, então o fato de que essas informações, que são pertinentes ao domínio e que devem ser mapeadas em sua ontologia, beneficiam a predição do unlink é exatamente a razão pela qual acreditamos na superioridade da predição semântica de unlinks em comparação às abordagens não semânticas.

3.2 Metodologia

Para realizar a predição dos unlinks de um estado para o outro da rede apresentamos o esquema da nossa proposta na Figura 12.

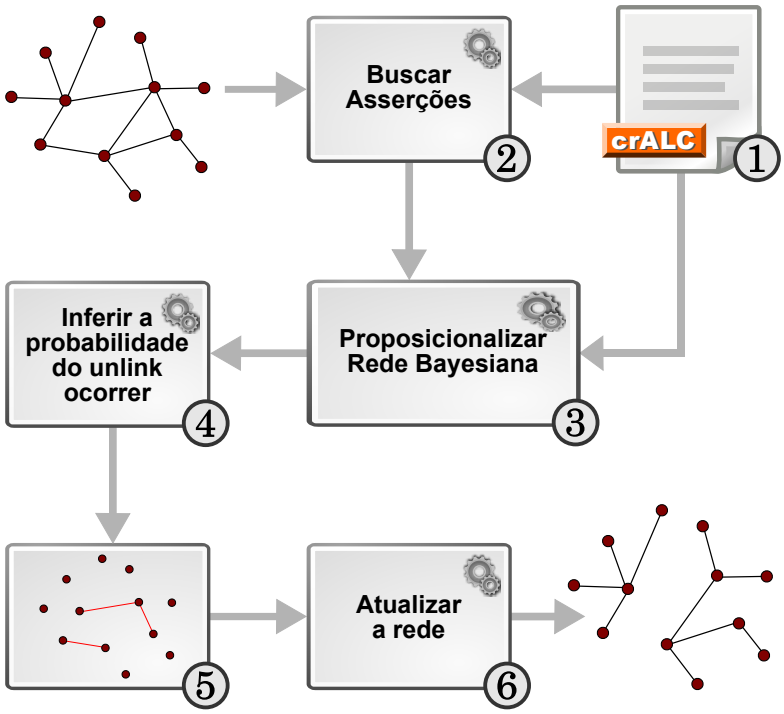
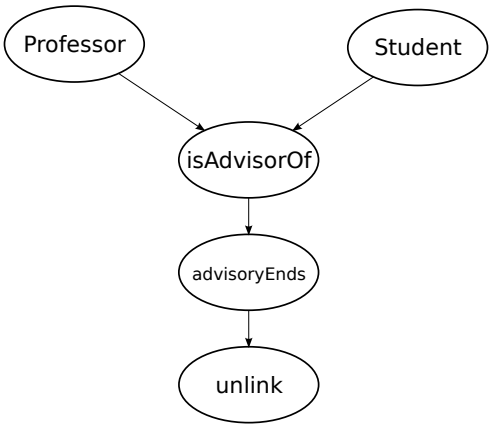


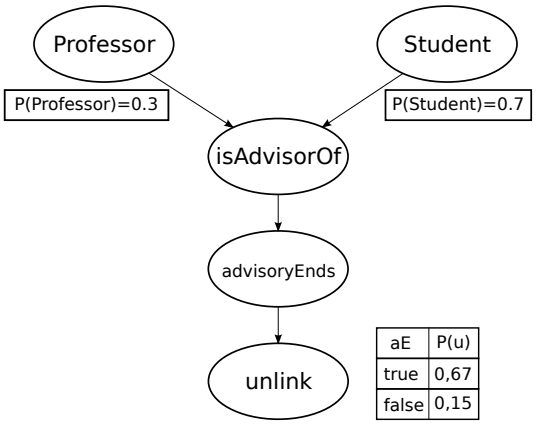
Figura 12: Proposta de solução para a predição de unlinks

No primeiro passo da figura 12, precisamos que uma ontologia probabilística referente ao domínio considerado e representada em CR$\mathcal{ALC}$ seja definida de forma manual ou automática [52]. Esta ontologia deve levar em consideração a tarefa de predizer o unlink, e por isso deve possuir um papel definindo a variável unlink(x,y) e os seus relacionamentos causais.

Após definida a ontologia, a distribuição das probabilidades condicionais pode ser realizada, estatisticamente, a partir das asserções existentes na base de conhecimento. A seguir, na Figura 13, apresentamos um exemplo levando em conta o domínio aluno-orientador. No nossa exemplo 30% dos indivíduos são Professores e os 70% são Alunos, além disso, das orientações finalizadas 67% dos casos resultaram em unlinks enquanto que das orientações que ainda estão em andamento apenas 15% resultaram em unlinks.



(a) sem a distribuição de probabilidade condicionais definidas



(b) com parte da distribuição de probabilidades condicionais definidas

Figura 13: Exemplo de uma ontologia Aluno-Orientador

Na etapa (2) da Figura 12, baseando-se na CR$\mathcal{ALC}$ definida e no par de nós do unlink que estamos considerando, são recuperadas todas as asserções relevantes ou um subconjunto dessas asserções que represente a realidade do conjunto total [11]. Elas são processadas e consolidadas em Aboxes, um para cada período de tempo. Uma rede bayesiana é (3) proposicionalizada, considerando as asserções do $Abox'_t(u, v)$ que é consolidado a partir das:

- asserções de todos os $Abox_i$ onde $0 < i < t + 1$ e $t + 1$ é o período de tempo em que estamos considerando se o unlink irá acontecer e
- e de novas asserções que sumarizem o conhecimento representado nestes Aboxes. Poderíamos, por exemplo, introduzir asserções que indiquem se o relacionamento entre dois pesquisadores é recente ou já vem de longa data. Por exemplo, se em 3 ABoxes consecutivos são verificadas asserções que indicam que dois pesquisadores mantiveram um relacionamento neste período de tempo então podemos adicionar uma nova asserção de $shortTermRelationship(u, v)$, que é um papel que representa a curta longevidade do relacionamento entre u e v, mas se o relacionamento entre estes dois pesquisadores ocorre em 5 ou 8 ABoxes consecutivos então, ao invés de $shortTermRelationship(u, v)$, podemos adicionar as asserção $mediumTermRelationship(u, v)$ ou

$longTermRelationship(u, v)$ respectivamente. As semânticas de *short, medium* e *long* são parâmetros da nossa proposta que foram definidos com os valores 3, 5 e 8 mas poderiam assumir outros valores. Outras asserções referentes a outros papéis poderiam ser adicionadas dependendo do domínio e da interpretação do conhecimento existente nos ABoxes disponíveis.

A rede bayesiana proposicionalizada é então utilizada para (4) inferir a probabilidade do unlink. As etapas 2, 3 e 4 da Figura 12 são repetidas para cada um dos nós conectados da rede.

Um ponto de corte é então utilizado para classificar os unlinks e aqueles que são identificados como prováveis unlinks são agrupados no conjunto E^{-}_{t+1} e finalmente a rede é atualizada removendo-se os unlinks que foram preditos, etapas (5) e (6) da Figura 12. O algoritmo da solução apresentada pode ser conferido a seguir no Algoritmo 1:

Algoritmo 1 Algoritmo de predição de unlinks

Require: uma rede G, um ontologia $\mathcal{O}$ e um ponto de corte γ.
Ensure: um conjunto de unlinks predito E^{-}_{t+1}

1: inicialize $E^{-}_{t+1} = \emptyset$;
2: **for all** triplas (u, v, t) em E_t **do**
3: consolida $\mathcal{A}'_t(u, v)$ a partir das asserções de todos $\mathcal{A}_i(u, v)$ onde 0<i<t+1 e as novas asserções que sumarizam o conhecimento representado nestes subconjuntos;
4: infere a probabilidade de $P(unlink(u, v)|\mathcal{A}'_t(u, v))$
usando a rede bayesiana criada a partir da ontologia $\mathcal{O}$;
5: **if** $P(unlink(u, v)|\mathcal{A}'_t(u, v)) > \gamma$ **then**
6: adiciona a tripla (u, v, t) em E^{-}_{t+1}
7: **end if**
8: **end for**

4 Experimento

Neste capítulo descreveremos o experimento executado para avaliar a nossa proposta de predição de unlinks descrevendo o cenário utilizado, a metodologia empregada, como executamos o experimento e quais os resultados obtidos comparando-os aos resultados de outros métodos de predição de unlinks.

4.1 Descrição do Cenário

Nossa solução utiliza uma rede evolutiva com as seguintes características:

- os relacionamentos entre os nós devem possuir a data em que aconteceram para que possamos criar subgrafos representando estados da rede em diferentes momentos do tempo.

- cada relacionamento existente na rede possui uma semântica, mas nessa mesma rede podemos ter relacionamento distintos com semânticas diferentes (multiplex). Normalmente nas amostras de redes disponibilizadas por outros pesquisadores na internet os relacionamentos possuem o mesmo sentido, provavelmente por conta do objetivo da pesquisa, mas isso não reflete a realidade da maioria das redes sociais. Numa rede social como o Facebook podemos encontrar diversos tipos de relacionamentos como "x é amigo de y", "x compartilhou a foto f", "x estudou na universidade u", "x leu o livro l", etc.

Não foi possível encontrar uma base de dados disponível que possuísse estas duas características. Optamos então por coletar nossa própria amostra de rede evolutiva baseada nos dados dos currículos disponibilizados pela Plataforma Lattes e pelo DBLP.

4.1.1 Dataset Lattes

A Plataforma Lattes[4] é o repositório público brasileiro de currículos científicos, contendo mais de 100.000 currículos de doutores e mais de 200.000 currículos de mestres

[4]http://lattes.cnpq.br/

nas mais variadas áreas de pesquisa. A informação disponibilizada pelo site é codificada em formato HTML, que além de dados pessoais como nome e endereço profissional inclui dados referentes às publicações, áreas de pesquisa, projetos e orientações de cada um dos pesquisadores.

Apesar da informação presente nestas páginas web seguir um modelo relacional (tabela1), uma rede evolutiva de colaboração pode ser construída através dos relacionamentos existentes e da identificação de novos relacionamentos implícitos nestes dados relacionais. Por exemplo, quando o pesquisador x informa os dados referentes a uma orientação, é possível identificar o relacionamento "x *orienta* y" onde x é o orientador e y é o pesquisador informado na coluna "aluno". Infelizmente, a coluna "aluno" não informa o identificador único do pesquisador na Plataforma Lattes mas simplesmente o nome por extenso do aluno. Precisamos então realizar a identificação pela comparação do nome do aluno com o nome dos pesquisadores na tabela "Pesquisadores" que é passível de erro pois pode levar à identificação de homônimos.

Em outros exemplos, podemos tentar identificar o relacionamento implícito "x *coautor* z" através da tabela "Publicações". Neste caso podemos realizar esta identificação de duas maneiras: a primeira seria encontrar dois registros de publicações, informados por pesquisadores distintos, com títulos idênticos no mesmo ano. Esta forma não é perfeita pois pode resultar na identificação errada de coautores pela possível existência de publicações homônimas. A segunda forma seria identificar os coautores através da coluna "autores" fornecida na tabela "Publicações" mas o conteúdo desta coluna é o nome abreviado, como utilizado em citações, dos autores da publicação que também pode levar a identificação de autores com nomes abreviados idênticos.

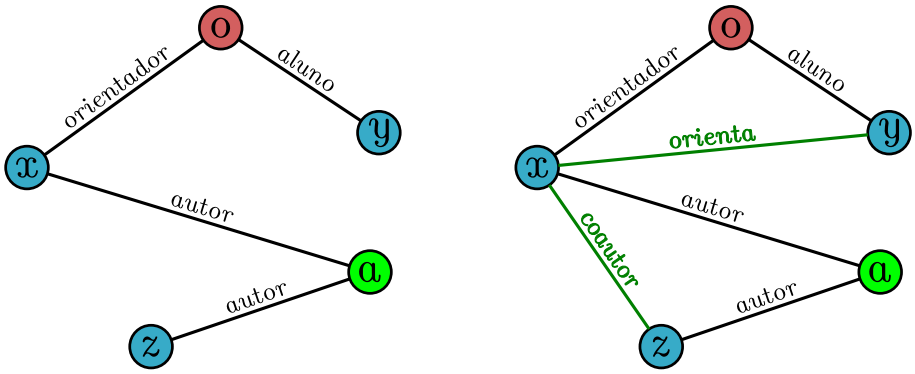


Figura 14: Identificando relacionamentos implícitos na Plataforma Lattes

Além dos problemas relacionados a homônimos, a identificação dos relacionamentos pela comparação de colunas textuais pode não acorrer caso os dados apresentem inconsistências, tais como, erros de digitação ou dados incorretos. Técnicas que tentam resolver estes problemas podem ser encontradas na literatura [53, 54, 55] mas estão fora do escopo deste trabalho.

Tabela 1: Tabelas e suas respectivas colunas no Dataset Lattes

Tabela	**Campos disponíveis**
Pesquisadores	nome, nomes utilizados em citações, endereço profissional
Publicações	pesquisador, título, autores, ano
Produções	pesquisador, título, autores, ano
Orientações	pesquisador, título, tipo, aluno, ano
Eventos	pesquisador, título, ano
Formações	pesquisador, tipo, ano início, ano conclusão, descrição
Colaboradores	pesquisador, colaborador
Áreas de pesquisa	pesquisador, área atuação,

O mesmo subconjunto de 8.000 pesquisadores, utilizado por Ochoa et al. [7], foi expandido e as informações referentes as publicações, produções, orientações, participação em eventos, formações realizadas, colaboradores e áreas de pesquisa de cada um destes pesquisadores foi recuperada. Os resultados desta etapa podem ser conferidos na Tabela 2.

Tabela 2: Quantidade de registros por tabela e links identificados no Dataset Lattes

Tabela	**Qtd. de Registros**
Pesquisadores	8.000
Publicações	1.064.280
Produções	94.247
Orientações	293.434
Eventos	252.691
Formações	35.295
Colaboradores	473.600
Áreas de pesquisa	30.068

(a) Tabelas

Link	**Qtd. de links identificados**
sharePublication	198.240
shareProduto	10.963
cooriented	2.724
isAdvisedBy	1.738
sameEvent	11.899
Total	225.564

(b) Links identificados

Figura 15: Representação da rede Lattes, exibindo apenas os nós com maior grau, as cores e o tamanho dos nós representam o grau do nó e as cores das arestas são as mesmas do nó de origem

4.1.2 Dataset DBLP

O Dataset DBLP[5], que foi utilizado em [37], é composto por artigos relacionados à ciência da computação provenientes de 28 conferências relacionadas a Mineração de Dados, Bancos de Dados e Aprendizado de máquina [37]. Cada um destes artigos, além de dados básicos como título e conferência, possui a lista dos seus autores. A partir destes dados foi possível construir um grafo de forma a representar a rede de coautoria dos artigos presentes no dataset, como ilustrado na Tabela 3. Configuração dos dados presentes no DBLP dataset. Neste grafo, os nós representam os pesquisadores e as arestas entre esses pesquisadores representam o relacionamento "é coautor de uma publicação com". Este relacionamento foi inferido a partir da seguinte regra: se dois pesquisadores r_1 e $r_2$são autores de uma publicação p, então r_1 é coautor de uma publicação com r_2. Todos os dados do dataset DBLP foram utilizados de forma a experimentar nossa solução em uma rede distinta e de proporções bem maior que a do Lattes.

Tabela 3: Quantidade de registros por tabela e links identificados no Dataset DBLP

Tabela	**Qtd. de Registros**
Researchers	1.253.089
Publications	2.337.573
Articles	555
Books	1.730
Proceedings	23.006
Wwws (sites)	869.149
Incollections	8.138
Inproceedings	1.434.995

(a) Tabela Researchers, tabela Publications e os tipos de publicações contidas em Publications

Link	**Qtd. de links identificados**
sharePublication	4.813.288
shareArticle	534
shareBook	2.007
shareWww	10.846
shareProceedings	6
shareIncollections	25.651
shareInproceedings	4.774.244

(b) Links sharePublication identificados e seus subtipos

[5]http://dblp.uni-trier.de/

O processo realizado para a montagem das redes é apresentado na Figura 16: (1) as informações são extraídas das redes sociais e armazenadas, em seguida (2) essas informações são convertidas em tabelas, a partir desses dados (3) os links e as datas das suas ocorrências são identificados. Finalmente, com os links identificados, é possível então (4) construir a rede evolutiva.

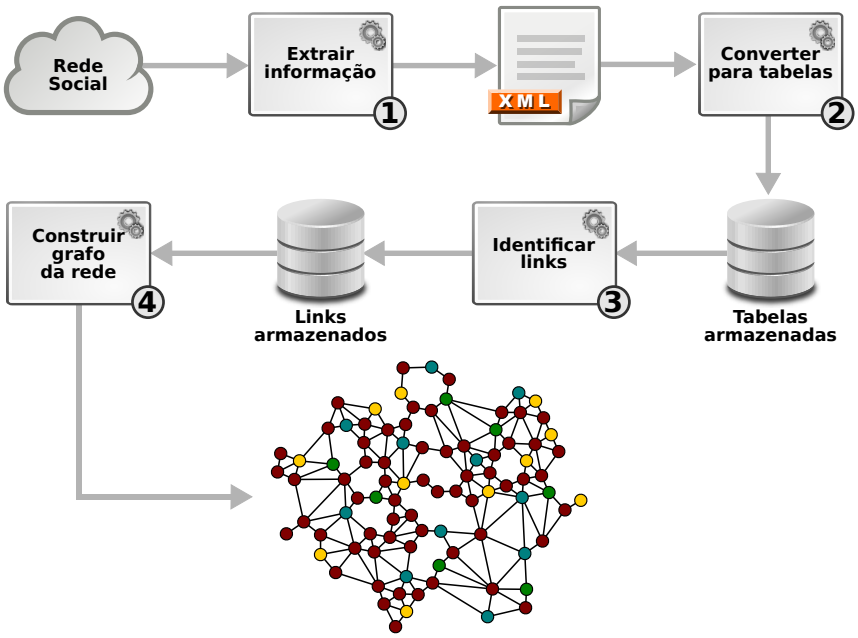


Figura 16: Montagem da rede

4.2 Metodologia Experimental

Para realizar o experimento de predição de unlinks, consideramos as colaborações entre pesquisadores encontradas nos datasets Lattes e DBLP. Da plataforma Lattes foram consideradas colaborações baseadas em relacionamentos de coautoria, coorientação e aluno-orientador entre os anos de 2006 à 2012. Os anos anteriores à 2006 foram desconsiderados pela baixa quantidade de links. Da plataforma DBLP foram consideradas colaborações baseadas apenas em relacionamentos de coautoria entre os anos 1992 à 2014. Os anos anteriores a 1992 foram desconsiderados pois a quantidade de links era baixa em comparação a quantidade de links de anos mais recentes como 2013. A evolução, dessas redes, pode ser conferida nas Tabelas 4 e 5:

Tabela 4: Rede Evolutiva Lattes (links ano a ano)

t	$t+1$	E_t	E_t^-	E_t^+	E_{t+1}
2000	2001	8	8	17	17
2001	2002	17	5	32	44
2002	2003	44	32	15	27
2003	2004	27	19	7	15
2004	2005	15	7	12	20
2005	2006	20	13	15101	15108
2006	2007	15108	11307	10612	14413
2007	2008	14413	10390	12932	16955
2008	2009	16955	12913	11024	15066
2009	2010	15066	11200	11121	14987
2010	2011	14987	11583	6813	10217
2011	2012	10217	7319	6483	9381

Tabela 5: Rede Evolutiva DBLP (links ano a ano)

t	$t+1$	E_t	E_t^-	E_t^+	E_{t+1}
1959	1960	65	65	1	1
1960	1961	1	1	18	18
1961	1962	18	17	113	114
1962	1963	114	113	2	3
1963	1964	3	3	13	13
1964	1965	13	12	46	47
1965	1966	47	43	199	203
1966	1967	203	201	279	281
1967	1968	281	270	452	463
1968	1969	463	460	357	360
1969	1970	360	356	310	314
1970	1971	314	286	687	715
1971	1972	715	703	401	413
1972	1973	413	395	614	632
1973	1974	632	576	1514	1570
1974	1975	1570	1514	880	936
1975	1976	936	882	1223	1277
1976	1977	1277	1199	1453	1531
1977	1978	1531	1475	1567	1623
1978	1979	1623	1468	1801	1956
1979	1980	1956	1782	2102	2276
1980	1981	2276	2065	2614	2825
1981	1982	2825	2561	3539	3803
1982	1983	3803	3472	3754	4085
1983	1984	4085	3770	4534	4849
1984	1985	4849	4482	4916	5283
1985	1986	5283	4769	7497	8011
1986	1987	8011	7283	7135	7863
1987	1988	7863	7043	9838	10658
1988	1989	10658	9637	11546	12567
1989	1990	12567	11226	15232	16573
1990	1991	16573	14620	16954	18907
1991	1992	18907	16243	20450	23114
1992	1993	23114	19437	30178	33855
1993	1994	33855	28909	35733	40679
1994	1995	40679	34841	37960	43798
1995	1996	43798	37329	39045	45514
1996	1997	45514	38011	49702	57205
1997	1998	57205	47831	56339	65713
1998	1999	65713	53967	67696	79442
1999	2000	79442	65163	77057	91336
2000	2001	91336	74295	91112	108153
2001	2002	108153	86923	113709	134939
2002	2003	134939	107520	140487	167906
2003	2004	167906	130255	186116	223767
2004	2005	223767	176944	203909	250732
2005	2006	250732	196435	234894	289191
2006	2007	289191	227595	270623	332219
2007	2008	332219	264993	294702	361928
2008	2009	361928	290449	330208	401687
2009	2010	401687	324582	343289	420394
2010	2011	420394	339522	375235	456107
2011	2012	456107	369002	396593	483698
2012	2013	483698	393510	395860	486048
2013	2014	486048	465634	68678	89092

4.2.1 Seleção dos casos para avaliação

O objetivo do experimento é verificar se nossa abordagem é capaz de classificar corretamente:

- os unlinks positivos: colaborações entre pares de pesquisadores que já existiam no intervalo de tempo anterior mas que terminam no atual. Item 1 da Figura 17;
- e os unlinks negativos: colaborações entre pares de pesquisadores que já existiam no intervalo de tempo anterior e que se renovam no atual. Item 2 da Figura 17.

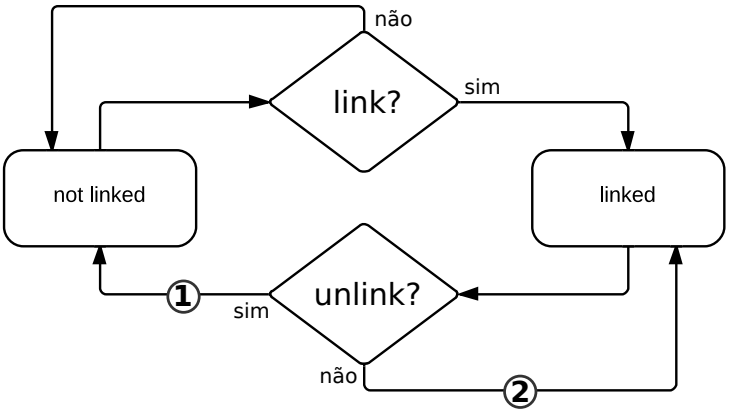


Figura 17: Unlink positivo (1) e unlink negativo (2)

Para realizar esta seleção estabelecemos as seguintes regras:

- $(u, v, t) \in E_t$

  O relacionamento entre dois pesquisadores se inicia em um algum momento do tempo t, sendo $2005 < t < 2012$ para o dataset Lattes e $1991 < t < 2014$ para o dataset DBLP.

- $(u, v, t + \alpha) \in E_{t+\alpha}$

  O relacionamento persiste por um tempo α, sendo $\alpha \geqslant 1$. Com essa regra selecionaremos colaborações que duraram pelo menos dois anos consecutivos. Utilizaremos o $(u, v, t + 1)$ como nosso unlink negativo.

- $(u, v, t + \alpha + 1) \notin E_{t+\alpha+1}$

  O relacionamento não se perpetua no ano seguinte. Utilizaremos o relacionamento $(u, v, t + \alpha + 1)$ como nosso unlink positivo.

- $t + \alpha + 1 < t_{last\_year}$

  No dataset Lattes, os relacionamentos referentes ao ano de 2012, ano em que os dados foram coletados, podem não ter sido totalmente recuperados. Optou-se então não considerar este ano, pois não temos como ter certeza de que os links existentes não tenham se perpetuado de 2011 para 2012, eles podem simplesmente não terem sidos observados no momento da criação do dataset, portanto definimos $t_{last\_year} = 2012$ para o dataset Lattes. O mesmo raciocínio foi utilizado para o dataset DBPL sendo neste caso $t_{last\_year} = 2013$.

De acordo com essas regras e com os Abox da Figura 18, apresentamos na Figura 19 alguns exemplos de colaborações selecionadas e outras desconsideradas. No exemplo (1) a colaboração entre Bob e Kim se inicia em 2007 e se mantém até 2010 mas não há registro de nenhum relacionamento entre os pesquisadores em 2011, quando então ocorre o unlink. No exemplo (2) a colaboração entre Bob e Dan se inicia sem 2006 mas não se renova em 2007 dessa forma não podemos utilizar este link para testar o unlink negativo.

$ABox_{2006}$	$ABox_{2007}$	$ABox_{2008}$	$ABox_{2009}$	$ABox_{2010}$	$ABox_{2011}$	$ABox_{2012}$
P(Bob)	P(Bob)	P(Bob)	P(Bob)	P(Bob)	P(Bob)	P(Bob)
	P(Kim)	P(Kim)	P(Kim)	P(Kim)		
P(Dan)						
		P(Ana)	P(Ana)	P(Ana)	P(Ana)	P(Ana)
P(Jen)	P(Jen)	P(Jen)	P(Jen)	P(Jen)	P(Jen)	
	l(Bob,Kim)	l(Bob,Kim)	l(Bob,Kim)	l(Bob,Kim)		
l(Bob,Dan)						
		l(Bob,Ana)	l(Bob,Ana)	l(Bob,Ana)	l(Bob,Ana)	l(Bob,Ana)
l(Bob,Jen)	l(Bob,Jen)	l(Bob,Jen)	l(Bob,Jen)	l(Bob,Jen)	l(Bob,Jen)	

Figura 18: ABoxes para o exemplo da aplicação das regras na seleção.
Onde: P(x) = Pesquisador(x) e l(x,y)=link(x,y)

No exemplo (3) o unlink não acontece dentro dos períodos observados, e portanto não há unlink positivo que possa ser usado para o experimento e esta colaboração é descartada. E finalmente no exemplo (4) a colaboração se mantém até 2011 mas não há registros do relacionamento em 2012. Como não temos certeza se neste último caso o relacionamento entre os pesquisadores realmente acabou ou se ele ainda não havia sido observado preferiu-se não utilizar esta colaboração.

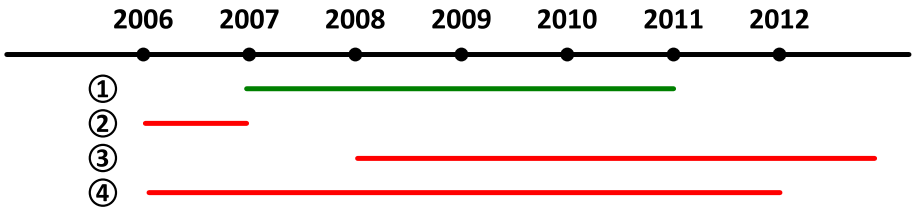


Figura 19: Exemplos de casos selecionados (1) e descartados (2,3,4)

Após análise dos dados do Lattes foram selecionadas 7918 colaborações que atendiam a todas as regras definidas. Desta forma temos 7918 casos de unlink positivos e 7918 casos de unlink negativos. No Dataset DBLP 50860 casos de unlink positivos e 50860 casos de unlink negativos foram selecionados.

4.2.2 Definição da CR$\mathcal{ALC}$

Para realizar a inferência da probabilidade do unlink entre dois indivíduos precisamos modelar o domínio através de uma ontologia probabilística, representada na lógica de descrição probabilística CR$\mathcal{ALC}$. Uma modelagem rica e significativa do domínio é importante mas, infelizmente, a maior parte das redes encontradas para a execução dos experimentos e avaliação dos resultados sofre pela falta de dados que expressem essa modelagem mais elaborada.

Mesmo o Dataset Lattes apesar de disponibilizar dados que possibilitam, por exemplo, a inclusão de conceitos como "isAdvisorOf" e "advisoryEnds" apresentados na Figura 11 não possui asserções suficientes destes conceitos para que a sua utilização na inferência seja significativa.

Para enriquecer a ontologia, propomos que a modelagem do domínio seja complementada com novos papéis que representem a temporalidade da rede. Para executar nossa proposta introduzimos três novos papéis (shortTermRelationship, mediumTermRelationship and longTermRelationship) que representam a longevidade do links no momento t de cada ABox. A semântica referente aos termos curto, médio e longo é obtida de acordo com parâmetros pré-definidos que podem, por exemplo, representar relacionamentos de 3, 5 e 8 anos respectivamente. Além destes introduzimos os papéis "growingNetwork" e "stableNetwork" seguindo as métricas propostas por [2].

Para facilitar a execução do experimento, evitando problemas de escalabilidade ou dificuldades na execução do experimento, modelamos a ontologia do domínio de tal forma a responder de forma similar a um classificador Naïve Bayes [56], demonstrando a versatilidade da nossa proposta. A ontologia utilizada é apresentada a seguir na Figura 20 e a sua representação em CR$\mathcal{ALC}$ utilizada neste trabalho pode ser conferida no Anexo I.

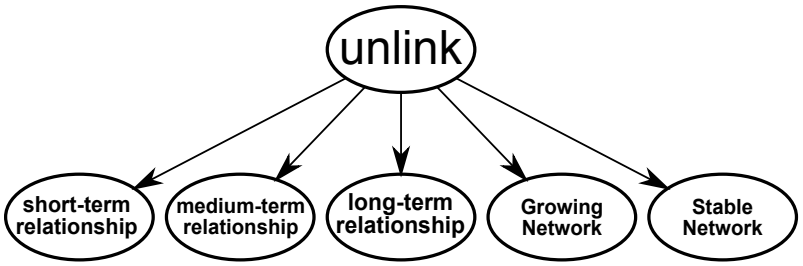


Figura 20: Naïve Bayes onde os atributos são roles e a classe é o role *unlink*

4.2.3 Execução do experimento

Para cada caso de unlink positivo e unlink negativo selecionado, através da metodologia explicada na seção 4.2.1, foram calculados os *scores* referentes a probabilidade de unlink utilizando os preditores baseados nas métricas de similaridade e o preditor baseado na CR$\mathcal{ALC}$ da nossa solução.

As métricas de similaridade, normalmente utilizadas para a predição de links, foram utilizadas neste trabalho de forma similar à aplicada em [10]. Para verificar sua utilidade na predição de unlinks, os autores definiram funções inversas dessas métricas: quanto maior o score obtido, que usualmente significaria uma maior probabilidade da predição positiva do link, menor é a probabilidade da predição do unlink. Na Figura 21 os procedimentos descritos neste paragrafo são representados pelos itens 1 e 2.

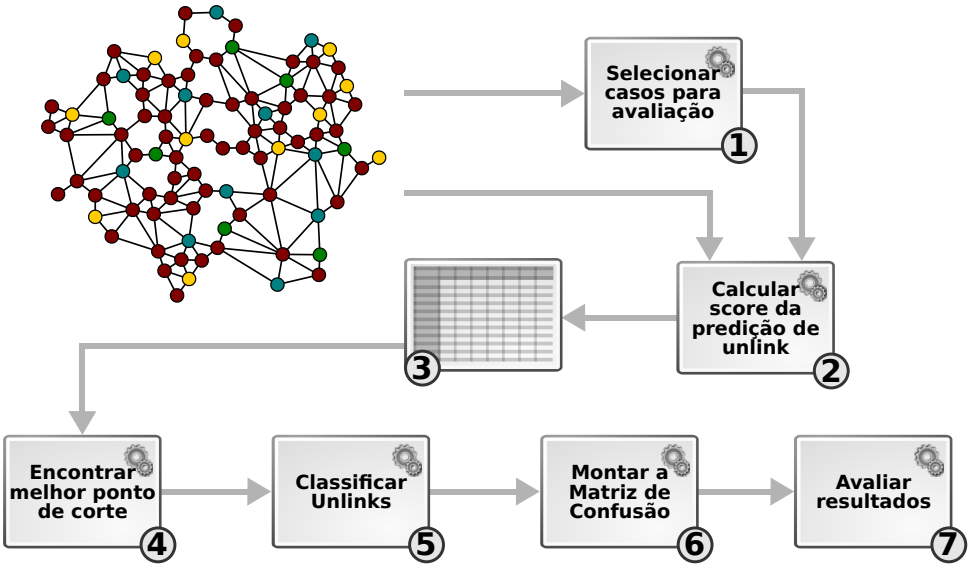


Figura 21: Diagrama da execução do experimento

Algoritmo 2 Execução do experimento

Require: A network G, a set of unlink predictors, a set of performance measures.
Ensure: a set o performance measure scores for each predictor

```
initialize Scores = ∅;
initialize Unlinks = ∅;
Unlinks = select triples (u,v,t) from G, representing unlinks (both postives and negatives), to be used in the unlink predictors evaluation;
for all Unlinks do
    for all predictors being tested do
        score(u,v,t) = predictor(G,u,v,t);
        add score(u,v,t) to Scores;
    end for
end for
for all predictors being tested do
    find predictor best threshold, the one that produces best accuracy
    create the confusion matrix by classifying the Unlinks using the defined threshold
    for all performance measures do
        calculate measure based on the defined confusion matrix
    end for
end for
```

Após o cálculo de todos os scores, o score que dividisse o conjunto de unlinks positivos e negativos com maior acurácia foi selecionado como ponto de corte. A partir deste pontos de corte, um para cada preditor, todos os casos foram classificados de acordo com os scores dos preditores. Na Figura 21 os procedimentos descritos neste paragrafo são representados pelos itens 4 e 5.

4.2.4 Avaliação

Por tratar-se de uma classificação binária, após a execução da predição de unlinks de todos os casos, é possível criar uma matriz de confusão para cada um dos preditores, como exemplificado na Figura 22, e a partir desta matriz avaliar a qualidade da classificação.

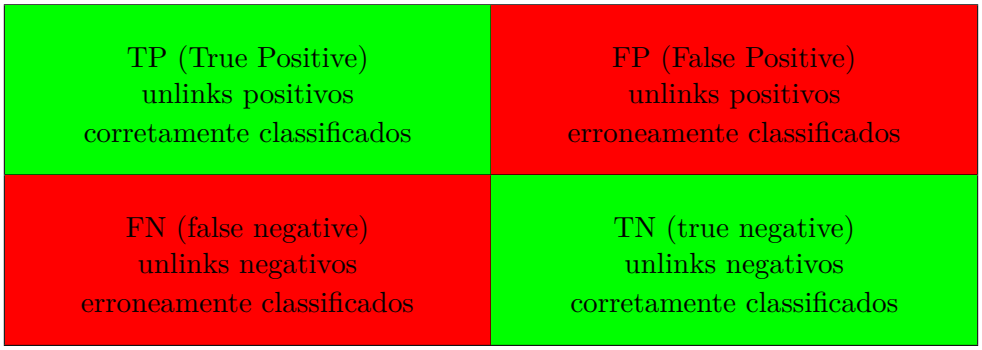


Figura 22: Matriz de confusão

Para avaliar a qualidade de cada um dos preditores foram utilizadas as seguintes medidas de avaliação por serem as mais utilizadas nesta tarefa [57, 13, 2, 58, 59]:

- **Precisão (Precision / Confidence)**

  É a fração dos unlinks positivos corretamente classificados pelo preditor em relação a todos os unlinks classificados como positivos.

$$Precision = PPV = \frac{TP}{TP + FP} \tag{16}$$

- **Recall (Cobertura / Sensitivity / True Positive Rate)**

  É a proporção dos unlinks positivos corretamente classificados pelo preditor em relação a todos os unlinks que são realmente positivos.

$$Recall = TPR = \frac{TP}{TP + FN} \tag{17}$$

A título de completitude apresentamos as outras :

- Specificity (True Negative Rate)

$$TNR = \frac{TN}{TN + FP} \tag{18}$$

- Fallout (False Positive Rate)

$$FPR = \frac{FP}{FP + TN} \tag{19}$$

- Miss Rate (False Negative Rate)

$$FNR = \frac{FN}{FN + TP} \tag{20}$$

- **Acurácia**

  É a proporção de unlinks corretamente classificados, sejam eles positivos ou negativos, em relação a todos os unlinks avaliados.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{21}$$

Entretanto estas medidas se analisadas isoladamente podem apresentar resultados enviesados quando as classes avaliadas estão desbalanceadas. Nestes casos a utilização de outras medidas é preferível.

- **F1 Score (Média Harmônica / F-Measure)**

  É a média harmônica entre a precisão e a cobertura. Sua definição completa é apresentada na Fórmula 22.

$$F_\beta = \frac{(\beta^2 + 1)(Precision \times Recall)}{\beta^2 Precision + Recall} \quad (0 \leq \beta \leq +\infty) \tag{22}$$

  β é o parâmetro que controla o balanço entre a precisão e o recall. Se $\beta > 1$, o valor do recall tem mais influência no cálculo da medida e se $\beta < 1$ o valor da precisão prevalece no cálculo. Quando $\beta = 1$ os valores de precisão e recall são considerados de forma balanceada no cálculo da medida, neste caso específico a medida também é conhecida como F1 score, pois $F_{\beta=1}$ [60]. A definição da F1 score é apresentada na Fórmula 23.

$$F_1 = \frac{2(Precision \times Recall)}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{23}$$

- **MCC (Matthews Correlation Coefficient)**

  O coeficiente de correlação de Matthews é uma aplicação do coeficiente de correlação de Pearson em matrizes de confusão [**58**]. Ele leva em conta os verdadeiros e falsos, positivos e negativos, e é geralmente considerado como uma medida equilibrada que pode ser utilizada mesmo em dados onde as classes são de tamanhos muito diferentes.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{24}$$

- **AUC (Area Under the Curve)**

  A curva ROC [61] é um gráfico plotado em duas dimensões onde o a taxa de TP (verdadeiros positivos) é representada no eixo Y e a taxa de FP (falsos positivos) é representada no eixo X [62]. Uma classificação totalmente aleatória traçaria uma linha reta que liga o origem (0,0) a (1, 1) enquanto que um classificador perfeito traçaria uma linha horizontal no topo do gráfico, como pode ser observado na Figura 23. O quanto uma curva ROC está mais próxima da linha perfeita e mais distante da linha diagonal aleatória define a qualidade do classificador.

  A área sob a curva ROC ou AUC (Area Under Curve) é uma medida padrão para a comparação de classificadores e pode ser obtida por métodos de integração numérica, como o método dos trapézios [62].

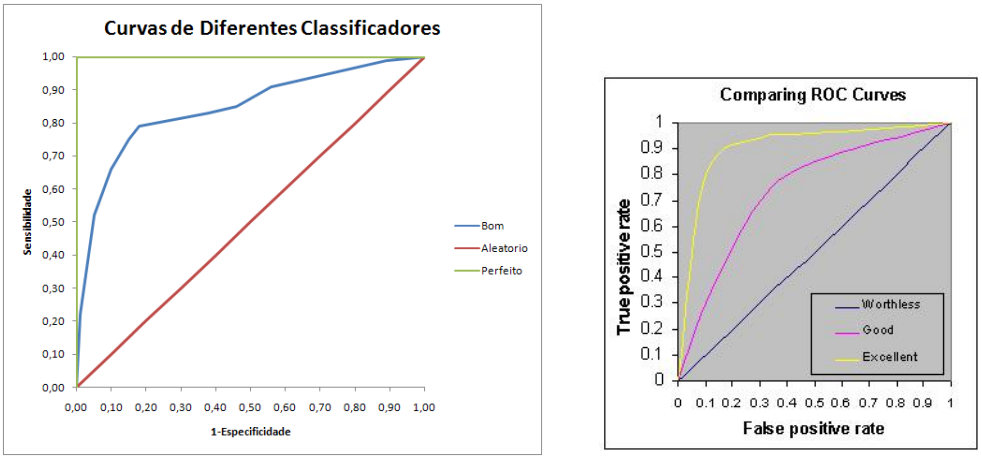


Figura 23: Curvas ROC de diferentes classificadores [3, 4]

4.3 Resultados

A seguir apresentamos na tabela 6 e nas figuras 24, 25 e 26 os resultados da avaliação dos preditores obtidos ao utilizarmos o dataset Lattes.

Tabela 6: Resultados da avaliação dos preditores (Lattes)

Preditor	Precisão	Recall	Acurácia	F1 Score	MCC	AUC
Common Neighbors	0.5537	0.5717	0.5555	0.5626	0.1111	0.5705
Jaccard Coefficient	0.5188	0.8830	0.5321	0.6536	0.0902	0.5323
Shortest Path	0.5188	**0.8833**	0.5321	**0.6537**	0.0903	0.5327
Pref. Attachment	0.5511	0.7279	**0.5675**	0.6273	**0.1426**	**0.5893**
Adamic/Adar	0.5350	0.6828	0.5447	0.5999	0.0930	0.5570
Katz	0.5471	0.6766	0.5582	0.6050	0.1199	0.5725
Event Based	0.5048	0.4624	0.5044	0.4827	0.0089	0.4926
Growth/Decay Based	**0.5768**	0.0810	0.5107	0.1421	0.0422	0.4226
Stability Based	**0.6025**	**0.8935**	**0.6520**	**0.7197**	**0.3473**	**0.7057**
CR$\mathcal{ALC}$Based	**0.6888**	**1.0000**	**0.7741**	**0.8157**	**0.6146**	**0.7628**

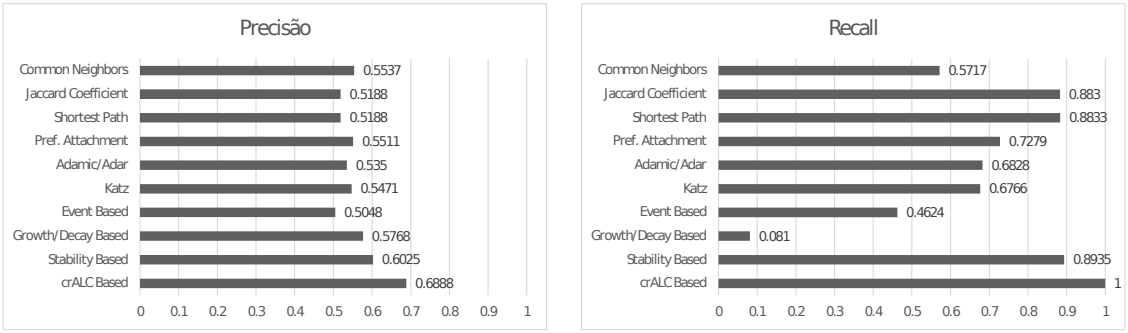


Figura 24: Resultados de Precisão e Recall para o Dataset Lattes

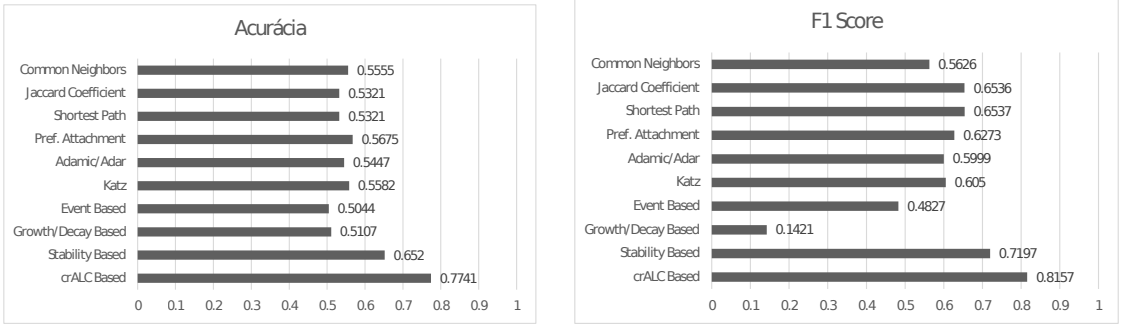


Figura 25: Resultados de acurácia e F1 Score para o Dataset Lattes

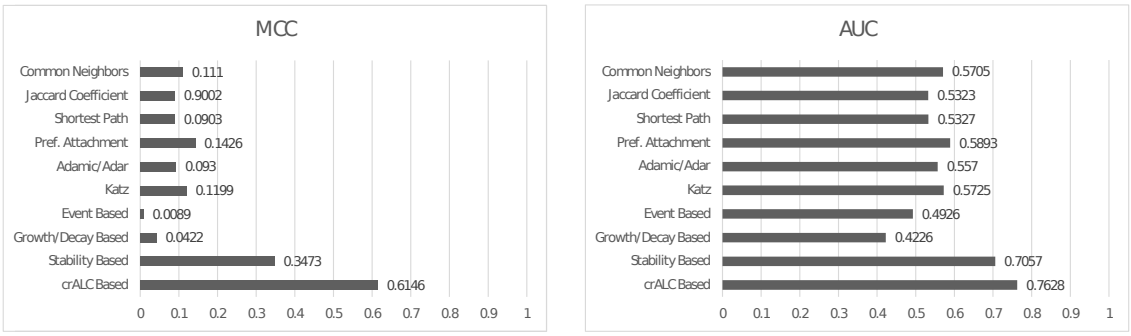


Figura 26: Resultados de MCC e AUC para o Dataset Lattes

Nossa proposta foi capaz de superar todas os outros preditores em todas as métricas utilizadas na avaliação. O preditor que conseguiu obter os resultados mais próximos foi o "Stability Based" um preditor que leva em consideração a temporalidade e que apresenta uma semântica que pode levar ao unlink: a instabilidade das conexões dos pesquisadores.

Outro preditor temporal que também possui semântica favorável ao unlink, o "Growth/Decay Based" que mede o nível de declínio da rede dos pesquisadores, apesar de apresentar boa precisão apresentou baixos resultados nas outras métricas de avaliação. Esse resultado talvez possa ser explicado pelo período curto dos dados, neste caso sete anos, que pode ser insuficiente para medir o crescimento ou declínio da rede dos pesquisadores analisados.

Os resultados obtidos pelo preditor "Event Based" foram baixos, mas provavelmente poderiam ser melhores com o ajuste fino dos valores configuráveis das recompensas para cada evento, merecendo ser avaliado novamente em experimentos futuros.

Os resultados referentes ao dataset DBLP, apresentados na tabela 7 e nas figuras 27, 28 e 29, foram ainda melhores que os apresentados no Lattes e mantém a superioridade da abordagem semântica.

Tabela 7: Resultados da avaliação dos preditores (DBLP)

Preditor	Precisão	Recall	Acurácia	F1 Score	MCC	AUC
Common Neighbors	**0.5667**	0.6460	**0.5761**	0.6038	**0.1537**	**0.5986**
Jaccard Coefficient	0.5230	0.7026	0.5309	0.5996	0.0658	0.5352
Shortest Path	0.5065	**0.9709**	0.5126	**0.6658**	0.0630	0.5125
Pref. Attachment	0.5607	0.6330	0.5685	0.5947	0.1383	0.5913
Adamic/Adar	**0.5754**	0.5121	0.5671	0.5419	0.1351	0.5815
Katz	0.5625	0.5763	0.5640	0.5693	0.1281	0.5868
Event Based	0.5067	0.3404	0.5045	0.4072	0.0096	0.5007
Growth/Decay Based	0.5246	0.4929	0.5231	0.5082	0.0463	0.5088
Stability Based	0.5666	**0.8108**	**0.5954**	**0.6671**	**0.2115**	**0.6303**
cr$\mathcal{ALC}$Based	**0.7324**	**1.0000**	**0.8173**	**0.8455**	**0.6818**	**0.8280**

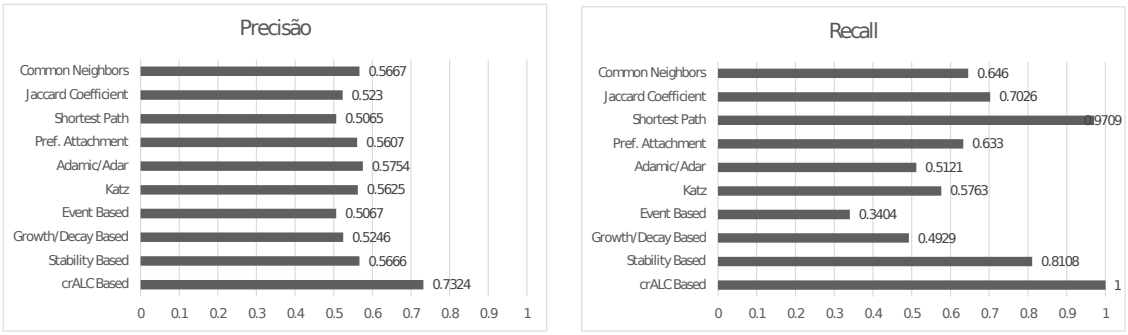


Figura 27: Resultados de Precisão e Recall para o Dataset DBLP

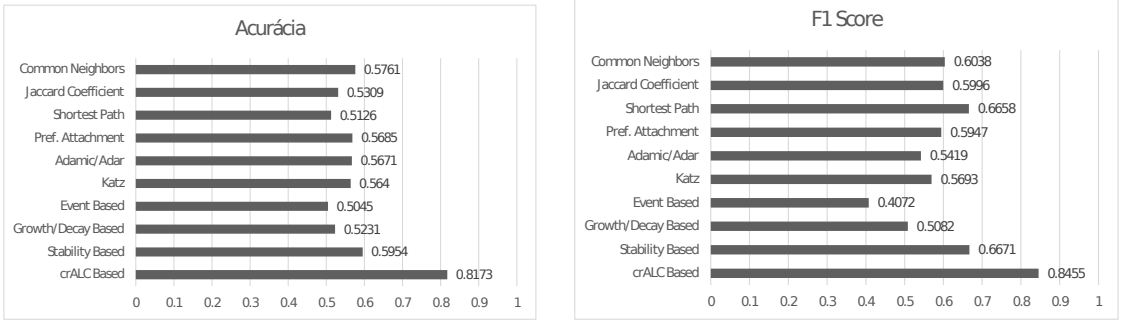


Figura 28: Resultados de Acurácia e F1 Score para o Dataset DBLP

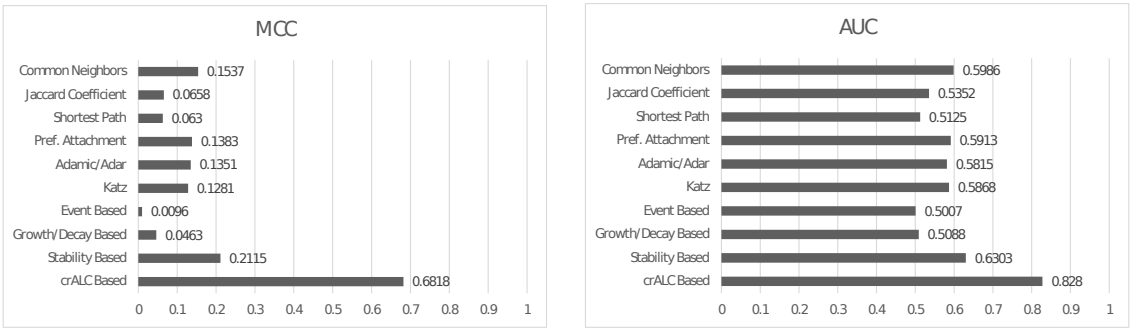


Figura 29: Resultados de MCC e AUC para o Dataset DBLP

4.4 Considerações finais

Neste trabalho buscamos avaliar se o conhecimento do domínio dos objetos da rede e os fatores temporais, mapeados através de uma ontologia probabilística, conseguiriam predizer os unlinks com boa qualidade. Conseguimos comprovar nossa hipótese ao aplicarmos nossa proposta no domínio de redes de coautoria. Além disso, ao predizer unlinks nas redes Lattes e DBLP nossa proposta conseguiu superar todas os outros abordagens avaliadas.

Os preditores baseados puramente em métricas estruturais e que não levam em consideração a temporalidade apresentaram resultados baixos. Os seus resultados demonstraram que eles não foram capazes de representar a semântica que leva ao unlink nas redes utilizadas no experimento.

A métrica que levou em consideração a estabilidade da rede dos pesquisadores apresentou bons resultados demonstrando que a importância dessa estabilidade na predição de unlinks em redes de coautoria.

5 Trabalhos Relacionados

Dada a importância do problema da predição de links é possível encontrar vários outros trabalhos que abordam este assunto. Em [38], os autores tentam resolver o problema da predição de links através do uso de "atributos de proximidade", "atributos de aglomeração" e "atributos topológicos".

Liben-Nowell e Kleinberg [6] acreditam que a evolução das redes sociais pode ser analisada de forma eficiente através de medidas de similaridade que se baseiam exclusivamente na topologia da rede mas citam que os motivos para o surgimento dos links muitas vezes transcendem as informações encontradas nas redes sociais, como quando um pesquisador se transfere para outro estado ou país e que essa nova proximidade geográfica influencia a formação de novos links. Seu trabalho contribuiu para a avaliação e comparação de várias métricas que vinham sendo apresentadas isoladamente em outros artigos. Apesar da contribuição significativa, os resultados obtidos não foram comparados a outros tipos de abordagens como a semântica.

O artigo de [63] apresenta a rede de coautoria Lattes discretizada em triênios, de forma que possamos analisar a evolução da rede, e o incremento da sua complexidade, de acordo com a dinâmica com que as novas colaborações surgem. Apesar de não ser um trabalho sobre predição de links, esta representação temporal poderia vir a ser utilizada em modelos de predição de links para prever a evolução da rede.

O fato de que os trabalhos anteriores em predição de links utilizam apenas o estado atual da rede para o cálculo da proximidade sem levar em consideração a informação temporal é destaque em [13]. Apesar de tratar-se de um trabalho sobre predição de links, a utilização da estrutura temporal das redes evolutivas bem como a métrica desenvolvida com base em eventos temporais são ideias aplicáveis e desejáveis na predição de unlinks. São citados outros trabalhos que utilizaram informações temporais mas nenhum referente a predição de unlinks.

Em [10] os autores tratam a questão da predição de unlinks como uma questão de identificar o enfraquecimento/decadência dos links. A abordagem estrutural foi escolhida para que o métodos utilizados pudessem ser utilizados nos mais diversos

tipos de redes independente do seu domínio. Os experimentos realizados utilizaram datasets representando a rede formada por artigos da Wikipedia e os links entre estes artigos.

Recentemente em [59] foi analisada extensivamente a forma como o problema da predição de links é avaliado na literatura: apontando problemas e levantando questionamentos sobre a seleção dos dados e métodos utilizados.

Tabela 8: Tabela comparativa dos trabalhos relacionados

Trabalho	Tipo de predição	Tipo de abordagem	Temporal?	Tipo de rede	Métricas utilizadas
Link prediction using supervised learning [38] Hasan, M. A.; Chaoji, V.; Salem, S. & Zaki, M. (2006)	Link	Estrutural	Não	Coautoria	- características de proximidade - algoritmos de classificação
The link-prediction problem for social networks [6] Liben-Nowell, D. & Kleinberg, J. M. (2007)	Link	Estrutural	Não	Coautoria	- Graph distance, common neighbors, Jaccard's coefficient, Adamic/Adar, preferential attachment, Katz, hitting time, commute time, rooted PageRank, SimRank
Proximity measures for link prediction based on temporal events [13] Soares, P. R. S. & Prudêncio, R. B. C. (2013)	Link	Estrutural	Sim	Coautoria	- Event-based, Time series, Adamic/Adar, common neighbors, Jaccard's coefficient, preferential attachment
DecLiNe-Models for Decay of Links in Networks [10] Preusse, J.; Kunegis, J.; Thimm, M. & Sisov, S. (2013)	Unlink	Estrutural	Não	Links entre páginas da Wikipédia	- Preferential attachment, Common neighbors, Cosine similarity, Jaccard index, Adamic–Adar
Predição de Links com CR$\mathcal{ALC}$	**Unlink**	**Semântica**	**Sim**	**Coautoria**	- Common Neighbors, Jaccard Coefficient, Shortest Path, Pref. Attachment, Adamic/Adar, Katz, Event Based, Growth/Decay Based, Stability Based, CR$\mathcal{ALC}$Based

6 Conclusão

6.1 Contribuições

Esta dissertação bem como os artigos originados da pesquisa para a realização deste trabalho são contribuições para área de predição de unlinks, área que ainda é pouco abordada pela comunidades acadêmica.

Avaliamos o desempenho de diversos métodos estruturais, largamente utilizados na predição de links, também para a tarefa de predição de unlinks.

Como contribuição tecnológicas estaremos disponibilizando um framework para avaliação de métodos de predição de links e unlinks capaz de:

6.1.1 Artigos publicados

Durante o desenvolvimento de todo o trabalho, o andamento da pesquisa gerou artigos que foram publicados em conferência, simpósio, e seminário.

No início da pesquisa para a definição do tema da dissertação, buscávamos ampliar e consolidar a predição de links utilizando a lógica de descrição probabilística CR$\mathcal{ALC}$ no artigo [64] foi apresentada a proposta inicial, o estudo de métricas de similaridade estruturais e semânticas que poderiam ser combinadas à CR$\mathcal{ALC}$ e a metodologia experimental que seria utilizada para avaliação da proposta. A partir dos resultados obtidos uma extensão do algoritmo de predição de links, mais eficiente e acurada, seria sugerida. Este artigo foi publicado e apresentado no Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI 2013) do IX Simpósio Brasileiro de Sistemas de Informação (SBSI 2013) realizado em João Pessoa - PB.

Posteriormente, numa fase onde a implementação do protótipo para a execução do experimento já era executável, a questão da escalabilidade levou a produção de um novo artigo [11]. O artigo discutia a importância da qualidade, ao invés da quantidade, das asserções utilizadas na inferência das probabilidades no resultado

do algoritmo de predição de links. O experimento realizado utilizava várias métricas diferentes para seleção de um conjunto de indivíduos pequeno, que forneceria um conjunto menor de asserções, mas que como verificado era capaz de obter o mesmo resultado que aqueles alcançados utilizando todos os indivíduos e suas asserções. Este artigo foi publicado e apresentado no 6º Seminário de Pesquisa em Ontologias do Brasil (ONTOBRAS 2013) realizado em Belo Horizonte.

Recentemente o foco da dissertação mudou da predição de links para a predição de unlinks. Ao analisarmos a evolução das redes sociais, a questão da predição de unlinks se apresentou como um problema interessante a ser resolvido e uma boa oportunidade para avaliar a predição utilizando CR$\mathcal{ALC}$ em outro contexto. Esta nova temática foi abordada no artigo "*Semantic Unlink Prediction in Evolving Social Networks through Probabilistic Description Logic*" [65] que é praticamente um resumo desta dissertação. Este artigo será publicado e apresentado no *Brazilian Conference on Intelligent System* (BRACIS 2014) que será realizado em outubro na cidade de São Carlos, SP.

6.1.2 Implementação

Foi implementado um pacote de software para a execução de todas as etapas do experimento.

- **Processamento dos arquivos XML referentes aos datasets Lattes e DBLP**

  Enquanto o arquivo XML contendo os dados do DBLP pode ser obtido diretamente no próprio site[6], o arquivo XML Lattes precisa ser extraído e compilado, para isso utilizamos o scriptLattes[66] ferramenta desenvolvida pelo CMCC-UFABC[7] e pelo CCSL-IME/USP[8]. Os arquivos então são processados e seus dados são armazenados em tabelas de um banco de dados relacional (PostgreSQL).

[6]http://dblp.uni-trier.de/xml/
[7]http://cmcc.ufabc.edu.br/
[8]http://ccsl.ime.usp.br/

- **Geração das redes evolutivas a partir dos dados processados dos datasets**

  A partir destes dados uma representação de rede evolutiva instanciada com a criação de um grafo da rede para cada intervalo de tempo. Para essa etapa utilizamos a biblioteca JGraphT[67]

- **Metodologia para a seleção dos casos que serão utilizados para a avaliação**

- **Predição de links e unlinks utilizando outras métricas de similaridade**

- **Predição de links e unlinks utilizando a lógica de descrição probabilística** CR$\mathcal{ALC}$

  Para isso o software deve interpretar a o arquivo com a especificação da CR$\mathcal{ALC}$ ler as asserções e através da proposicionalização gerar uma rede bayesiana. A inferência dessa rede bayesiana pode ser executada utilizando três motores diferentes:

  - JavaBayes[68]
  - SMILE[69]
  - AC-FOVE[70]

- **Implementação de métricas para a avaliação dos resultados**

  Realizar a avaliação utilizando as métricas de avaliação implementadas e novas que venham a ser adicionadas.

6.2 Limitações

A principal limitação do trabalho desenvolvido é a ausência de datasets disponíveis com semântica suficiente para a criação de ontologias mais complexas. A criação de datasets mais elaborados a partir do cruzamento de dados entre redes onlines e

a captura destes dados não é possível devido aos termos de uso impostos por estes sites:

- Facebook[9]: "3.2 You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission".
- ResearchGate[10]: "5.1.2 Users must not misuse the Service. Misuse of the Service includes, without limitation: automated or massive manual retrieval of other Users' profile data ("data harvesting");"

Outra dificuldade se deve ao processo de proposicionalização da rede bayesiana, etapa explicada na seção 2.1.3, que dependendo da CR$\mathcal{ALC}$ elaborada e da quantidade de indivíduos presentes na base de conhecimento pode inviabilizar a geração da rede bayesiana e a inferência das variáveis desejadas. Para evitar qualquer problema deste tipo, buscamos evitar construtores de restrição universal ($\forall r.C$) neste momento da pesquisa onde o foco é avaliar os benefícios da predição semântica de unlinks.

6.3 Trabalhos Futuros

Nossa proposta precisa ser avaliada em redes geradas a partir de bases de conhecimento que permitam a utilização de ontologias mais complexas. Para isso serão necessário novos datasets que já apresentem tais condições ou datasets que possibilitem o enriquecimento através da identificação de relacionamentos implícitos ao domínio ou identificados por especialistas. O enriquecimento também poderia ser concretizado através da execução de pesquisas à internet e outras redes sociais utilizando palavras-chave como, por exemplo, o nome do pesquisador.

A evolução das técnicas de alinhamento de ontologias[71] traz a perspectiva de integração de bases de conhecimento que antes eram desconexas e que passam a

[9]https://www.facebook.com/legal/terms
[10]http://www.researchgate.net/application.TermsAndConditions.html

contribuir com novos dados e novas semânticas que podem ser utilizados para a predição de links de unlinks.

Nossa proposta deve ser avaliada em novos domínios que não apenas o domínio de redes acadêmicas de co-autoria para que seja verificada sua aplicabilidade nesses novos domínios e se os resultados são semelhantes aos obtidos até agora.

Outra extensão deste trabalho seria predizer links e unlinks em conjunto, e passar a avaliar a predição da evolução da rede como um todo. Além disso, seria interessante resgatar trabalhos passados e incluir no processamento a metodologia de seleção de indivíduos apresentada em [11] e outras métricas semânticas estudadas em [64].

As probabilidades podem vir a ser definidas no modelo automaticamente através das asserções existentes na base de conhecimento, poderíamos também definir um conjunto de probabilidades para cada intervalo do tempo e probabilidades para intervalos futuros poderiam ser definidas através de regressão linear.

Como estamos trabalhando com redes complexas e pensando num futuro onde cada vez mais teremos redes se interligando com outras redes, poderíamos ter probabilidades diferentes para diferentes agrupamentos dentro dessas grandes redes.

Referências

[1] Renata Vieira, Débora Abdalla Santos, Douglas Michaelsen Silva, and Menandro Ribeiro Santana. Web semântica: ontologias, lógica de descrição e inferência. In *Web e Multimidia: Desafios e Soluções (WebMedia 2005-Minicursos)*, pages 127–167. 2005.

[2] Julia Preusse, Matthias Thimm, Thomas Gottron, and Steffen Staab. Structural Dynamics of Knowledge Networks. *Association for the Advancement of Artificial Intelligence (www.aaai.org)*, pages 506–515, 2013.

[3] César Souza. Análise de Poder Discriminativo Através de Curvas ROC. Disponível em: <http://crsouza.blogspot.com.br/2009/07/analise-de-poder-discriminativo-atraves.html> Acesso em 27 Agosto 2014.

[4] Thomas G. Tape. The Area Under an ROC Curve. Disponível em: <http://gim.unmc.edu/dxtests/roc3.htm> Acesso em 27 Agosto 2014.

[5] Mohammad Al Hasan and Mohammed J. Zaki. A Survey Of Link Prediction In Social Networks. In *Social network data analytics*, pages 243–275. 2011.

[6] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[7] José Eduardo Ochoa-Luna, Kate Revoredo, and Fabio Gagliardi Cozman. Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, pages 1–13, 2013.

[8] Pierre Borgnat, Eric Fleury, Jean-Loup Guillaume, Clémence Magnien, Céline Robardet, and Antoine Scherrer. Evolving networks. In *Proceedings of NATO Advanced Study Institute on Mining Massive Data Sets for Security*. IOS Press, 2008.

[9] Réka Albert and Albert-László Barabási. Topology of Evolving Networks: Local Events and Universality. *Physical Review Letters*, 85(24):5234–5237, December 2000.

[10] Julia Preusse, Jérôme Kunegis, Matthias Thimm, and Sergej Sizov. DecLiNe-Models for Decay of Links in Networks. *arXiv preprint arXiv:1403.4415*, 2014.

[11] Marcius Armada, Kate Revoredo, José Eduardo Ochoa Luna, and Fabio Gagliardi Cozman. Assertion Role in a Hybrid Link Prediction Approach through Probabilistic Ontology. *ONTOBRAS 2013*, pages 106–117, 2013.

[12] F. G. Cozman and R. B. Polastro. Complexity analysis and variational inference for interpretation-based probabilistic description logics. In *Conference on Uncertainty in Artificial Intelligence*, 2009.

[13] P.R.S. Soares and R.B.C. Prudêncio. Proximity measures for link prediction based on temporal events. *Expert Systems with Applications*, 40(16):6652–6660, 2013.

[14] F. Baader and W. Nutt. Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press, 2002.

[15] Manfred Schmidt-Schauß and Gert Smolka. Attributive concept descriptions with complements. *Artificial intelligence*, 48(1):1–26, 1991.

[16] I Horrocks. Ontologies and the semantic web. *Communications of the ACM*, 51:58–67, 2008.

[17] *MONK: Proposta de um Motor de InferÊncia Híbrido para a Web Smântica*. PhD thesis, Universidade Federal de Pernambuco, 2010.

[18] Rodrigo Bellizia Polastro. *Lógica Probabilística Baseada em Redes Bayesianas Relacionais com Inferência em Primeira Ordem*. PhD thesis, University of São Paulo, 2012.

[19] F Baader, D Calvanese, D McGuinness, D Nardi, and P Patel-Schneider. The Description Logic Handbook: Theory, Implementation, and Applications. *Cambridge University Press, New York*, 2002.

[20] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, 1990.

[21] Lise Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. Learning probabilistic models of relational structure. *ICML*, 2001.

[22] Brian Milch and Stuart J. Russell. First-order probabilistic languages: Into the unknown. In Stephen Muggleton, Ramón P. Otero, and Alireza Tamaddoni-Nezhad, editors, *ILP*, volume 4455 of *Lecture Notes in Computer Science*, pages 10–24. Springer, 2006.

[23] Fabio Gagliardi Cozman. Credal networks. *Artif. Intell.*, 120(2):199–233, 2000.

[24] Ronald Fagin, Joseph Y. Halpern, and Nimrod Megiddo. A logic for reasoning about probabilities. *Information and Computation*, 87:78–128, 1990.

[25] M. Jaeger. Relational bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6, 2002.

[26] Aline D Bessa, Leonardo B L Santos, Lorena P N R Martinez, Mariana C Costa, and Pedro G S Cardoso. Introdução às Redes Complexas. Technical report, 2010.

[27] Mark Newman. *Networks: An Introduction*. 2010.

[28] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8. 1994.

[29] P Erdös and A Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[30] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[31] M. Newman. The structure and function of complex networks. 2003.

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

[33] Stanley Milgram. The Small-World Problem. *Psychology Today*, 1:61–67, 1967.

[34] M.S. Granovetter. The strength of weak ties. *American Journal of Sociology*, (78):1360–1380, 1973.

[35] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, 2005.

[36] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey, 2011.

[37] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 322–331, Washington, DC, USA, 2007. IEEE Computer Society.

[38] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. 2006.

[39] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models, 2009. cite arxiv:0912.5410Comment: 96 pages, 14 figures, 333 references.

[40] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[41] Ben Taskar, Ming-fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances*

in Neural Information Processing Systems (NIPS) 16. Cambridge, MA: MIT Press, 2004.

[42] Waleed Aljandal, Vikas Bahirwani, Doina Caragea, and William H. Hsu. Ontology-aware classification and association rule mining for interest and link prediction in social networks. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 3–8. AAAI, 2009.

[43] Doina Caragea, Vikas Bahirwani, Waleed Aljandal, and William H. Hsu. Ontology-based link prediction in the livejournal social network. In Vadim Bulitko and J. Christopher Beck, editors, *SARA*. AAAI, 2009.

[44] Andreas Thor, Philip Anderson, Louiqa Raschid, Saket Navlakha, Barna Saha, Samir Khuller, and Xiao-Ning Zhang. Link prediction for annotation graphs using graph summarization. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist, editors, *International Semantic Web Conference (1)*, volume 7031 of *Lecture Notes in Computer Science*, pages 714–729. Springer, 2011.

[45] M E Newman and M E Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98:404–9, 2001.

[46] Gerard Salton and Michael J McGill. *Introduction to Modern Information Retrieval*. 1986.

[47] David Liben-Nowell, Jon M. Kleinberg, and Jon Kleinberg David Liben-nowell. The Link Prediction Problem for Social Networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2004.

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

[49] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[50] MA Brandão and MM Moro. Using link semantics to recommend collaborations in academic social networks. *Proceedings of the 22nd …*, (Section 3):833–840, 2013.

[51] Xiongcai Cai, Michael Bain, and Alfred Krzywicki. Learning collaborative filtering and its application to people to people recommendation in social networks. *Data Mining (ICDM), …*, 2010.

[52] José Eduardo Ochoa Luna, Kate Revoredo, and Fabio Gagliardi Cozman. Learning probabilistic description logics: A framework and algorithms. In *MICAI (1)*, pages 28–39, 2011.

[53] Xiaoming Fan, Jianyong Wang, Xu Pu, Lizhu Zhou, and Bing Lv. On Graph-Based Name Disambiguation. *Journal of Data and Information Quality*, 2(2):1–23, February 2011.

[54] Hai Jin, Li Huang, and Pingpeng Yuan. Name disambiguation using semantic association clustering. In *Proceedings - IEEE International Conference on e-Business Engineering, ICEBE 2009; IEEE Int. Workshops - AiR 2009; SOAIC 2009; SOKMBI 2009; ASOC 2009*, pages 42–48, 2009.

[55] Dongwook Shin, Taehwan Kim, Hana Jung, and Joongmin Choi. Automatic Method for Author Name Disambiguation Using Social Networks. *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 1263–1270, 2010.

[56] Mohammed J. Zaki and Wagner Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[57] Yugchhaya Dhote, Nishchol Mishra, and Sanjeev Sharma. Survey and analysis of temporal link prediction in online social networks. *Advances in Computing,*

Communications and Informatics (ICACCI), 2013 International Conference on, pages 1178–1183, 2013.

[58] DMW Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, (December), 2011.

[59] Ryan Lichtnwalter, N.V. Nitesh V. Chawla, and R. Lichtenwalter. Link prediction: Fair and effective evaluation. volume 0 of *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pages 376–383, Los Alamitos, CA, USA, 2012. IEEE Computer Society.

[60] Yutaka Sasaki. The Truth of the F-measure, 2007. Disponível em: <http://www.cs.odu.edu/ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf> Acesso em 27 Agosto 2014.

[61] Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. In *NIPS*, 2004.

[62] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical report, HP Laboratories, Palo Alto, 2003.

[63] Jesús Mena-Chalco, Luciano Digiampietri, and Roberto Cesar Jr. Caracterizando as redes de coautoria de currículos lattes. In *CSBC 2012 - BraSNAM ()*, jul 2012.

[64] Marcius Armada and Kate Revoredo. Predição Semântica de Links : algoritmos e aplicações. 2013.

[65] Marcius Armada and Kate Revoredo. Semantic Unlink Prediction in Evolving Social Networks through Probabilistic Description Logic.

[66] Jesús Pascual Mena-Chalco and Roberto Marcondes Cesar Junior. scriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39, December 2009.

[67] Barak Naveh and John Sichi. JGraphT, 2013.

[68] Fabio Gagliardi Cozman. Generalizing Variable Elimination in Bayesian Networks. In *Workshop on Probabilistic Reasoning in Artificial Intelligence at SBIA/Iberamia*, pages 21 – 26, 2000.

[69] MJ Druzdzel. SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: a development environment for graphical decision-theoretic models. *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pages 902–903, 1999.

[70] Felipe I. Takiyama and Fabio G. Cozman. Inferences on CRALC Using the AC-FOVE Algorithm, 2012.

[71] Alex Alves, Anselmo Guedes, Kate Revoredo, and Fernanda Baião. Classificador de alinhamento de ontologias utilizando técnicas de aprendizado de máquina. *IX Simpósio Brasileiro de Sistemas de Informação*, pages 25–36, 2013.

Anexo I - CR$\mathcal{ALC}$s utilizadas nos experimentos

Com o intuito de permitir a reprodutibilidade do experimento apresentado nesta dissertação, apresentamos a seguir as CR$\mathcal{ALC}$s que foram utilizadas pelo nosso algoritmo de predição de unlinks.

(probability unlink(x,y) 0.5)
(conditional-probability isNetGrowing(x,y) unlink(x,y) 0.38987 0.38179)
(conditional-probability isNetStable(x,y) unlink(x,y) 0.47436 0.52564)
(conditional-probability oneYearLongLink(x,y) unlink(x,y) 0.64713 0.19058)
(conditional-probability twoYearLongLink(x,y) unlink(x,y) 0.19058 0.08676)
(conditional-probability threeYearLongLink(x,y) unlink(x,y) 0.08676 0.049)
(conditional-probability fourYearLongLink(x,y) unlink(x,y) 0.049 0.02652)
(conditional-probability fiveYearLongLink(x,y) unlink(x,y) 0.02652 0.0)
(conditional-probability hasEventRelationship(x,y) unlink(x,y) 0.13008 0.10697)
(conditional-probability hasCoorientationRelationship(x,y) unlink(x,y) 0.0581 0.04193)
(conditional-probability hasProductionRelationship(x,y) unlink(x,y) 0.10773 0.08664)
(conditional-probability hasPublicationRelationship(x,y) unlink(x,y) 0.12099 0.87901)

Figura 30: CR$\mathcal{ALC}$utilizada para a predição de unlinks no Dataset Lattes

(probability disconnected(x,y) 0.5)
(conditional-probability isNetGrowing(x,y) disconnected(x,y) 0.49291 0.44665
(conditional-probability isNetStable(x,y) disconnected(x,y) 0.85598 0.66545)
(conditional-probability oneYearLongLink(x,y) disconnected(x,y) 0.0 0.63472)
(conditional-probability twoYearLongLink(x,y) disconnected(x,y) 0.63472 0.21150)
(conditional-probability threeYearLongLink(x,y) disconnected(x,y) 0.21150 0.08230)
(conditional-probability fourYearLongLink(x,y) disconnected(x,y) 0.082309 0.08230)
(conditional-probability fiveYearLongLink(x,y) disconnected(x,y) 0.07145 0.07145)

Figura 31: CR$\mathcal{ALC}$utilizada para a predição de unlinks no Dataset DBLP