Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP

Luciano Digiampietri[1], **Ricardo Linden**[2], **Lenin Barbosa**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH-USP)

[2]Faculdade Salesiana Maria Auxiliadora (FSMA)

digiampietri@usp.br, rlinden@pobox.com, lenin.ferreira@gmail.com

Abstract. *Author name identification is fundamental to calculate correctly bibliometric metrics, but homonyms and polyseme come in the way of correct identification, making it necessary to apply an author name disambiguation algorithm. This paper presents a disambiguation technique that consists on automatic feature extraction followed by the application of a classifier. The results obtained in the case study achieved 96% precision, which is very similar to the state of the art in the literature.*

Resumo. *A identificação de autores é fundamental a precisão dos cálculos bibliométricos, mas homônimos e polissemia atrapalham a identificação, sendo necessária a aplicação de um algoritmo de desambiguação de nomes de autor. Este artigo propõe uma técnica de desambiguação que consiste em extração automática de características seguidas da aplicação de um classificador. Os resultados obtidos no estudo de caso atingiram precisão de 96%, similar ao estado da arte da literatura.*

1. Introdução

Para a melhor distribuição de recursos é necessário medir a produtividade das pessoas, o que, no Brasil é feito, em grande parte, pela produção bibliográfica de cada professor/cientista. Para isto, precisamos descobrir exatamente quem é o autor de cada trabalho, o que pode ser complicado devido à existência de polissemia (várias maneiras diferentes de escrever o mesmo nome) e homonímia (nomes diferentes que são escritos da mesma forma e pessoas diferentes com exatamente o mesmo nome). Um exemplo disto é a busca de GUPTA, R. no DBLP[1], onde encontramos 113 registros de autores diferentes, o que ilustra a homonímia. Por outro lado, se buscarmos no mesmo DBLP o nome RAJESH GUPTA, verificamos que existem dois possíveis registros para seu nome (Rajesh K. Gupta e Rajesh Gupta). Em ambos os casos, se não soubermos separar aqueles trabalhos que realmente desejamos, podemos obter métricas errôneas para a produtividade de cada autor (muito maior para o caso de GUPTA, R. e menor para o caso de RAJESH GUPTA). Assim, precisamos de uma forma de retirar esta ambiguidade dos nomes de autores.

Esta necessidade gerou toda uma área de pesquisa denominada *Author Name Disambiguation* (AND), ou desambiguação de nome de autor, que tem sido um alvo daqueles pesquisadores interessados em cientometria e análise de redes sociais acadêmicas. Vários métodos avançados têm sido usados para tentar resolver este problema de forma

[1]dblp.uni-trier.de

efetiva e neste artigo propomos a aplicação de um método de seleção de atributos combinado com um classificador para obter uma solução de qualidade para este problema.

2. Trabalhos Relacionados

Primeiramente, temos que ver as informações que são usadas para a desambiguação. Ferreia et al. (2012) descrevem quatro principais informações utilizadas na desambiguação a partir de referências bibliográficas. A informação primordial é a lista de autores, com as limitações decorrentes de grafia errada e/ou incompleta de nomes e registros de nomes idênticos para autores distintos. Podemos usar também as outras informações das citações, como ano, título do veículo e do artigo, permitindo fazer verificações sobre a viabilidade da autoria de um pesquisador específico. Também podemos usar a informação da rede social de coautorias, analisando, entre outras coisas, se dois autores da rede de mesmo nome possuem vizinhos (coautores) em comum, o que pode indicar tratar-se da mesma pessoa. Por fim, pode-se buscar na web dados adicionais sobre os autores (por exemplo, instituição onde trabalha, áreas de atuação e orientações) a fim de utilizar essas informações no processo de desambiguação.

Smalheiser and Torvik (2009) explicam porque não podemos confiar em um registro central de identificadores unívocos de autores, como o ORCID, pois precisaríamos de uma participação maciça, inclusive retroativa. Ademais, os autores também apontam que desambiguação manual tende a obter resultados diferentes de acordo com os realizadores. Assim, evidencia-se a necessidade de se buscar os dados nas próprias referências de forma automática.

Strotmann e Zhao (2012) mostram que técnicas de AND eliminam os multiautores (agrupamentos de autores de nomes similares), que são ocorrências comuns, especialmente devido à comunalidade de certos sobrenomes orientais. Assim, aplicações de cientometria que usam AND obterão resultados mais precisos.

Milojevic (2013) argumenta que, na maioria das vezes, métodos simples baseados no nome e iniciais dos autores (o primeiro tipo de informações descrito na seção anterior) são suficientes para identificar a maioria dos autores (até 97%, de acordo com o estudo), em bases com poucos multiautores ou em situações que não envolvam busca de extremos (autores mais produtivos, por exemplo). Ainda assim, grande parte dos artigos buscam algoritmos complexos para resolver este problema. Exemplificando com referências altamente citadas, temos que Han et al (2004) usam máquina de vetores de suporte e métodos naïve Bayes. enquanto que Song et al (2007) usam dois modelos hieráquicos Bayesianos para realizar a desambiguação em um grande conjunto de citações do Citeseer. Entretanto, o resultado colocado por Milojevic sugere que classificadores baseados em técnicas mais simples como o *Rotation Forest* (baseado em árvores de decisão) podem ter um resultado comparável com aqueles mais complexos.

Existem várias aplicações práticas para desambiguação de nomes. Liu et al (2014), por exemplo, criaram um *framework* de agrupamento com várias métricas e o aplicaram ao *PubMed*[2] com bons resultados. Ademais, a desambiguação é um passo necessário em várias aplicações de nível mais alto, como aquela em cientometria descrita por Lima et al (2015), que visa a avaliar a produtividade dos autores. Assim, a proposição de um método que obtenha resultados melhores é algo de grande interesse.

[2]http://www.ncbi.nlm.nih.gov/pubmed

3. Metodologia

A metodologia deste trabalho foi organizada nas seguintes atividades: revisão da literatura correlata; obtenção dos dados; seleção da amostra; anotação manual da amostra; extração das características; teste, validação e análise dos resultados.

Obtenção dos dados: inicialmente foram baixadas as listas dos professores permanentes dos programas de pós-graduação em Computação no Brasil da CAPES[3] (que serão utilizados para a seleção da amostra). Adicionalmente, foram copiados todos os dados do projeto DBLP (*Digital Bibliography & Library Project*), em novembro de 2014, consistindo em milhões de autores e de publicações. Cada autor possui uma chave de identificação, mas podem existir falhas na atribuição de chaves, como vimos nos casos identificados na introdução deste artigo. Assim, é possível que existam diferentes registros que se referem a um mesmo autor e, assim, no presente artigo objetiva-se aprimorar o processo de resolução de nomes de autores nos dados do DBLP.

Seleção da amostra: optou-se utilizar como amostra os registros de autores do DBLP que possuíssem os primeiros e últimos nomes compatíveis como uma lista de nomes pré-selecionados (os nomes dos 48 professores permanentes do programa de pós-graduação em Ciência da Computação da UNICAMP, um dos programas nota 7 da área de Ciência da Computação). Selecionamos, então, no conjunto de registros de autores do DBLP todos aqueles que possuíssem o primeiro nome e o último nome (sobrenome) dos professores da lista de 48 nomes utilizada, criando-se blocos com pessoas que possuíssem esses dois nomes em comum. Cada par de autores de cada bloco foi avaliada como sendo ou não a mesma pessoa. Este processo resultou na seleção de 82 registros de autores do DBLP, com a identificação única e correta de 29 professores e a criação de 17 blocos ambíguos contendo entre 2 e 12 registros de autores cada (na média, cada bloco possui aproximadamente três registros de autores)[4].

Anotação manual da amostra: cada par de registros de autores de cada bloco (totalizando 102 pares) foi anotado manualmente para identificação se correspondiam ou não à mesma pessoa para os treinamentos e testes da solução proposta. Este processo foi realizado por duas pessoas, a primeira fazendo a anotação e a segunda a validação.

Extração das características: quatorze características de quatro tipos foram extraídas para cada um dos 102 pares de nomes pertencentes ao mesmo bloco (Tabela 1). As informações utilizadas para a extração das características são: nomes dos autores; rede social de coautorias (esta rede foi gerada a partir de todas as publicações de artigos em periódicos e em anais de eventos disponibilizadas pelo DBLP, cerca de 2,8 milhões de publicações); mineração de texto baseada nos títulos dos artigos publicados pelos autores da amostra (1.828 artigos); e período no qual cada autor realizou suas publicações (informação também extraída dos 1.828 artigos encontrados).

Teste, validação e análise dos resultados: as estratégias usadas para analisar a importância de cada característica, bem como para avaliar a acurácia da combinação das características extraídas a partir dos dados serão detalhadas na próxima seção.

---

[3] http://www.capes.gov.br/component/content/article?id=4656:ciencia-da-computacao

[4] Dois professores não tiveram nenhum registro encontrado no DBLP.

Tabela 1. Características extraídas das citações

Tipo de característica	Característica	Descrição
características da rede social de coautorias	vizinhos em comum	número de vizinhos em comum na rede social acadêmica de todos os autores da DBLP
	são vizinhos	indica se o par de autores é ou não vizinho na rede social
características extraídas dos nomes	distância de edição	distância de edição entre os nomes dos dois autores
	distância relativa	proporção entre a distância de edição e o tamanho do menor nome dentre os autores
	primeiro nome diferente	indica se os autores têm seus primeiros nomes diferentes (um do outro)
	último nome diferente	indica se os autores têm seus últimos nomes diferentes (um do outro)
	proporção de diferentes nomes do meio	proporção (em relação ao número total de nomes) de nomes diferentes entre os autores, contabilizando nomes adicionais como diferentes
	proporção de diferentes abreviações	proporção (em relação ao número total de nomes) de nomes abreviados diferentes entre os autores, contabilizando abreviações adicionais como diferentes
	nomes invertidos	indica se há inversão da posição das partes dos nomes entre os autores
	nomes ou abreviações diferentes	indica a proporção de nomes efetivamente diferentes entre os autores (sem contar a presença de nomes adicionais)
características baseadas na mineração de texto	mineração de texto dos títulos dos artigos	métrica baseada em TFIDF que compara a frequência das palavras dos títulos entre os dois autores e entre o corpus formado pelos títulos de todos os autores avaliados
	log(MT)	logaritmo do valor resultante da mineração de texto
características baseadas nos anos de publicação dos artigos	intersecção do período de publicação	interseção entre os períodos de publicação dos dois autores
	distância em anos entre publicações	distância mínima em anos entre as publicações dos dois autores (apenas se a interseção for igual a zero)

4. Resultados

As características extraídas para cada um dos 102 pares de registros de autores utilizados na amostra foram inicialmente estudadas em relação às suas contribuições para a desambiguação de nomes. A Figura 1 apresenta a correlação de Pearson entre a classe e as demais características[5]. O atributo *classe* indica se o par de autores corresponde a uma só pessoa. Pode-se ver que as três maiores correlações (todas negativas) com a *classe* ocorrem entre atributos relacionados aos nomes dos autores *proporção de diferentes nomes do meio*, *nomes ou abreviações diferentes* e *último nome diferente*, evidenciado o fato esperado de que as diferenças nos nomes dos autores são as principais características básicas para a desambiguação. Porém, os desafios da desambiguação de nomes ocorre justamente quando a comparação entre nomes por si só não consegue evidenciar se os nomes pertencem ou não à mesma pessoa. A característica não relacionada aos nomes que obteve a maior correlação com a *classe* é *vizinhos em comum* (correlação igual a 0,32), indicando que para nomes iguais ou compatíveis a presença de vizinhos em comum na rede de coautorias é um indício importante de que os dois nomes se referem à mesma pessoa. Também obtiveram correlações positivas os atributos *log(MT)* e *intersecção do período de publicação*. Em especial, destaca-se que a mineração de textos realizada sobre os títulos, apesar de ser bastante simplista, auxilia no processo de desambiguação.

[5]A característica "são vizinhos" indica que os dois nomes aparecem em posições diferentes da lista de coautores de ao menos uma publicação. Como isto nunca ocorreu para a amostra analisada, esta característica foi descartada do trabalho e não aparece na figura

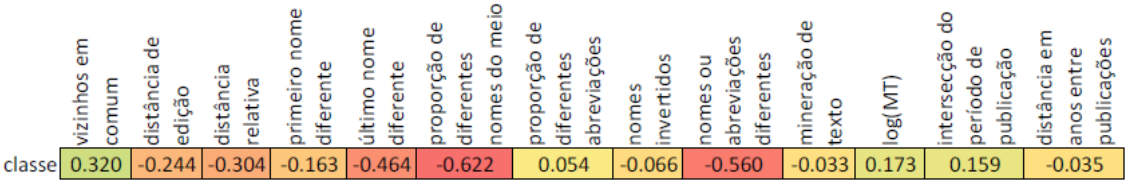

Figura 1. Correlação entre as características e a classe

Foram utilizados diferentes seletores de atributos[6] para identificar quais atributos são mais relevantes para em relação ao atributo *classe*. Os atributos mais selecionados são: *proporção de diferentes nomes do meio*, *nomes ou abreviações diferentes*, *último nome diferente*, *mineração de texto*, *log(MT)* e *primeiro nome diferente*. Além das características extraídas de nomes, foram selecionadas apenas as características extraídas a partir de mineração de textos. Destaca-se que não se observou nos trabalhos correlatos o uso da mineração de textos aplicada a títulos como usado no presente trabalho.

O próximo passo consistiu em fazer o problema de desambiguação ser tratado como um problema de classificação. Agora, cada par de autores que potencialmente correspondem a mesma pessoa deveria ser classificado pelo algoritmo como verdadeiro (são a mesma pessoa) ou falso (são pessoas diferentes). Para tanto, as características selecionadas foram combinadas utilizando-se o metaclassificador *Rotation Forest* [Rodriguez et al. 2006] e o desempenho da solução foi avaliado utilizando-se a validação cruzada em 10 subconjuntos. As medidas de despenho utilizadas são: Taxa de Verdadeiro-Positivos (VP) para cada classe; Taxa de Falso-Positivos (FP); Precisão; Revocação; Medida-F; e Área ROC. Estas medidas são apresentadas na Tabela 2.

Observa-se que a taxa de Verdadeiro-Positivos para a classe F (F), isto é, pares que não correspondem à mesma pessoa, foi de 1 (100%), ou seja, todos os pares deste tipo foram classificados corretamente como não sendo a mesma pessoa. Já esta métrica para a classe T (T) foi de 66,7% (foram identificados corretamente dois terços dos pares que correspondem à mesma pessoa). A média ponderada para esta métrica foi de 96,1%.

Tabela 2. Desempenho da solução proposta

classe	VP	FP	Precisão	Revocação	Medida-F	Área ROC
F	1	0,333	0,957	1	0,978	0,977
T	0,667	0	1	0,667	0,8	0,977
Média Ponderada	**0,961**	**0,294**	**0,962**	**0,961**	**0,957**	**0,977**

Os quatro erros de classificação ocorreram quando pares de nomes referentes à mesma pessoa foram identificados como sendo de pessoas diferentes. Nestes quatro casos um dos nomes do par a ser avaliado era composto por apenas dois nomes (primeiro nome e último sobrenome) sem evidências adicionais de que são a mesma pessoa.

5. Conclusões e Trabalhos Futuros

Neste trabalho foram pesquisadas e analisadas diferentes características que podem ser extraídas de dados de referências de publicações a fim de se desambiguar os nomes dos autores. Quatorze características foram extraídas dos dados do DBLP e suas importâncias na

[6]Foram utilizados os seletores de atributos do arcabouço Weka: http://www.cs.waikato.ac.nz/ml/weka/

agregação de informação sobre os autores foram analisadas considerando-se a correlação de Pearson e diferentes seletores de atributos. Um metaclassificador foi utilizado para combinar as características extraídas de forma a classificar pares de nomes de autores como *pertencentes à mesma pessoa* ou *não pertencentes à mesma pessoa*. Os resultados da classificação, para a amostra selecionada, atingiram uma precisão média de 96% e uma medida-F superior a 0,95.

Apesar dos resultados iniciais serem bastante promissores, destacamos que a amostra utilizada era pequena (com apenas 12 instâncias da classe V). Assim, precisamos de mais testes para termos mais segurança quanto à eficácia e à eficiência da estratégia utilizada. Ademais, era uma base que não foi usada em outros trabalhos, o que dificulta a comparação com os resultados obtidos por outros autores.

Como trabalhos futuros pretende-se melhorar o processo de seleção de atributos para aumentar a quantidade de informação fornecida para o algoritmo, executar testes utilizando bases de outros autores, amostras bem maiores e incluir novas características a fim de tornar a estratégia de desambiguação mais robusta.

Referências

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, 41(2):15–26.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Chen, H., editor, *4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 296–305.

Lima, H., Silva, T., Moro, M., Santos, R., Meira, Wagner, J., and Laender, A. (2015). Assessing the profile of top Brazilian computer science researchers. *Scientometrics*, pages 1–18.

Liu, W., Dogan, R. I., Kim, S., Comeau, D. C., Kim, W., Yeganova, Z., and Lu, Z. (2014). Author name disambiguation for pubmed. *Journal of the Association for Information Science and Technology*, 65(4):765–781.

Milojevic, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, 7(2):767–773.

Rodriguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630.

Smalheiser, N. R. and Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, 43(1):1–43.

Song, Y., Huang, J., Council, I. G., Li, J., and Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In Sugimoto, S., editor, *JCDL '07 Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 342–351.

Strotmann, A. and Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 63(9):1820–1833.