em questão
E-ISSN 1808-5245

Desambiguação de nomes de autores para a identificação automática de perfis acadêmicos

Luciano Antonio Digiampietri

Doutor; Universidade de São Paulo, São Paulo, SP, Brasil;
digiampietri@usp.br

João Eduardo Ferreira

Doutor; Universidade de São Paulo, São Paulo, SP, Brasil;
jef@ime.usp.br

Resumo: A desambiguação de nomes é uma atividade fundamental em estudos bibliométricos, em particular naqueles que utilizam diferentes fontes de informação. O objetivo deste trabalho é propor e testar uma estratégia de desambiguação de nomes de autores de forma a possibilitar a identificação automática do perfil do Google Acadêmico de docentes. A estratégia proposta é baseada na busca pelos perfis dos docentes no Google Acadêmico, seguida por um processo de casamento de nomes. Adicionalmente são comparadas as publicações acadêmicas que estão cadastradas no currículo Lattes do docente e no perfil do Google Acadêmico. Por fim, a resolução de nomes ocorre, verificando-se entre os perfis compatíveis aquele que apresenta maiores evidências de pertencer ao respectivo docente. Um estudo de caso envolvendo os docentes da Universidade de São Paulo foi realizado, e o sistema automático foi capaz de identificar, de maneira correta, 4.283 perfis do Google Acadêmico. Uma análise de cobertura mostrou que o sistema foi capaz de encontrar cerca de 95% dos perfis dos docentes que possuem essa informação, e nenhum falso-positivo foi identificado.

Palavras-chave: Desambiguação de nomes. Resolução de entidades. Bibliometria.

1 Introdução

A atividade de desambiguação de nomes é uma atividade bastante desafiadora, pois mesmo quando todas as bases ou registros envolvidos têm o nome completo da entidade, existem entidades diferentes com o mesmo nome (homônimas), e o nome de entidades pode mudar ao longo do tempo. Uma pessoa, ao se casar, por exemplo, pode incorporar o sobrenome do cônjuge ao seu.

A desambiguação de nomes de autores (*author name disambiguation* – AND) é uma atividade extremamente importante para estudos bibliométricos. Esta atividade visa identificar e agrupar todos os registros de um dado autor e pode ser utilizada para, por exemplo, integrar diferentes bases de dados acadêmicas, ou para agrupar os registros de um dado autor em uma base que não esteja padronizada. Um dos grandes desafios da AND é que, tipicamente, o nome dos autores em registros acadêmicos não está completo. Em dados relacionados a referências bibliográficas, é comum a presença apenas do primeiro nome e do sobrenome, ou ainda do sobrenome e da inicial do primeiro nome.

Para enfrentar alguns dos desafios da desambiguação de nomes de autores é comum utilizar informações complementares, além do nome dos autores. Entre as informações mais utilizadas estão o título das publicações, o ano das publicações, os veículos nos quais as publicações foram realizadas e informações sobre a vinculação do autor.

O trabalho visa apresentar uma estratégia para a desambiguação de nomes de autores que considera informações de três bases de dados acadêmicas: a base de dados institucional dos docentes da Universidade de São Paulo (USP), os currículos da Plataforma Lattes e os perfis de autores do Google Acadêmico. O objetivo do trabalho é identificar a correta ligação entre os docentes da USP, cujos currículos Lattes já estavam previamente identificados, com o registro correspondente de cada docente no Google Acadêmico.

O restante do artigo está organizado da seguinte forma: a seção 2 apresenta os conceitos básicos e os trabalhos correlatos. A seção 3 contém a descrição dos materiais e métodos utilizados. Já a seção 4 apresenta a análise dos resultados obtidos. Por fim, a seção 5 contém as considerações finais e os trabalhos futuros.

2 Conceitos básicos e trabalhos correlatos

A expressão "resolução de entidades" (*entity resolution*) corresponde ao processo de determinar se duas referências a objetos do mundo real referem-se ou não ao mesmo objeto (TALBURT, 2010). No contexto acadêmico, a resolução de entidades costuma ser utilizada para identificar se (TALBURT, 2010; BORGES

et al., 2011; CANUTO et al., 2013; DIGIAMPIETRI, BARBOSA; LINDEN, 2015):

a) diferentes referências a publicações se referem à mesma publicação;
b) se referências a pessoas referem-se a uma mesma pessoa;
c) se referências a instituições referem-se a uma mesma instituição;
d) se diferentes referências a veículos de publicação correspondem ao mesmo veículo.

Talburt (2010) dividiu o processo de resolução de entidades em cinco atividades:

a) extração de referências a entidades - identificação e obtenção de dados referentes à entidades;
b) preparação das referências das entidades - limpeza e estruturação das informações das referências;
c) resolução das referências a entidades - determinação se duas referências referem-se ou não à mesma entidade;
d) gerenciamento das entidades identificadas - criação e manutenção de uma base de entidades;
e) análise do relacionamento das entidades - análise dos relacionamentos existentes entre as diferentes entidades.

A atividade de resolução de entidades, propriamente dita, tipicamente utiliza medidas de distância entre os nomes das referências, mas pode, também, considerar informações adicionais. Por exemplo, ao se comparar duas referências a publicações é comum realizar a comparação dos títulos e, também, de informações adicionais como o ano da publicação, a lista de autores e o veículo no qual o trabalho foi publicado.

O processo de resolução de entidades aplicado à resolução de nomes de autores é chamado de desambiguação do nome do autor (*author name disambiguation* – AND). Esse processo é fundamental em estudos bibliométricos e em análise de redes sociais acadêmicas, servindo para evitar dois problemas: a atribuição incorreta de trabalhos a um dado autor e a não atribuição de um trabalho

ao seu autor quando, por exemplo, o nome na referência não é exatamente igual ao nome completo do autor.

Os nomes dos autores utilizados em referências bibliográficas podem apresentar dois problemas principais. O primeiro é a polissemia, isto é, um mesmo autor tem seu nome representado de diferentes formas nas referências bibliográficas. O outro problema é a homonímia, isto é, diferentes autores têm seus nomes escritos da mesma forma nas referências bibliográficas.

Apesar dos desafios da resolução de nomes de autores, Strotmann, Zhao (2012) e Milojevic (2013) afirmam que, ao se analisar pequenos grupos de entidades (por exemplo, os docentes de um departamento e suas produções bibliográficas), técnicas bastante simples de resolução de nomes podem atingir resultados bastante satisfatórios. Eles citam que apenas a verificação da inicial do primeiro nome e o sobrenome do autor pode atingir acurácias de até 97% na resolução de nomes. No entanto, ao se considerar grupos maiores, com centenas ou milhares de entidades, o resultado desse tipo de técnica é bastante inferior (STROTMANN; ZHAO, 2012; MILOJEVIC, 2013). Há outras estratégias que utilizam apenas o nome dos autores de maneira um pouco mais complexa, considerando, por exemplo, todas as partes dos nomes disponíveis e levando em conta a presença de erros de digitação. Tais estratégias também conseguem atingir acurácias bastante elevadas, ao tratar grupos pequenos de autores ou fontes de dados, nas quais se sabe de antemão quais são os autores que poderão estar presentes em cada referência (MUGNAINI et al., 2012).

Gomide, Kling e Figueiredo (2017) identificaram e analisaram os diferentes padrões nos usos dos nomes dos autores em citações, destacando que a análise da estrutura da rede de coautorias e os padrões de uso dos nomes podem ser utilizados para o desenvolvimento de algoritmos mais eficientes de desambiguação.

Técnicas mais sofisticadas aplicadas à resolução de nomes de autores, ou de publicações utilizam a análise de cocitação (WHITE; MCCAIN, 1998). Adicionalmente, diferentes tipos de medidas podem ser calculadas e, posteriormente, combinadas para identificar se duas referências a autores referem-se à mesma pessoa. Por exemplo, é possível comparar os assuntos das

publicações das respectivas referências, o período em que os autores realizaram suas publicações, as palavras-chave utilizadas, etc. Essas diferentes medidas podem ser combinadas utilizando-se, por exemplo, modelos bayesianos, máquina de vetores de suporte e redes neurais artificiais (HAN et al., 2004; SONG et al., 2007; DIAZ-VALENZUELA; MARTÍN-BAUTISTA; VILA, 2014; DIGIAMPIETRI; BARBOSA; LINDEN, 2015). Atualmente, além da combinação dessas diferentes medidas, alguns métodos de resolução de nomes também utilizam o *feedback* do usuário para melhorar o processo de resolução (FERREIRA; MACHADO; GONÇALVES, 2012; GODOI et al., 2013).

O trabalho utiliza, além do nome dos autores, as informações sobre suas publicações. A combinação dos resultados do casamento entre nomes de autores e do casamento entre a lista de publicações é feita por um conjunto de regras baseadas em termos linguísticos, atribuídos ao resultado do casamento (como *igual* ou *compatível*), conforme será apresentado na próxima seção. Este tipo de solução segue os princípios da lógica difusa (NOVÁK; PERFILIEVA; MOCKOR, 2012).

Ferreira, Gonçalves e Laender (2012) propuseram uma taxonomia para descrever as estratégias utilizadas na desambiguação de nomes de autores (figura 1). A taxonomia proposta organiza as estratégias em duas características principais: o tipo de estratégia e as evidências exploradas.

Figura 1 - Taxonomia para a classificação dos métodos de desambiguação do nome de autores.

Fonte: Traduzido de Ferreira, Gonçalves e Laender (2012).

O trabalho utiliza a estratégia de classificação, isto é, dada uma referência; no caso, um perfil do Google Acadêmico, o sistema verifica se ela é compatível com algum dos docentes da USP analisados e, se sim, faz a classificação, ou, especificamente, a atribuição do perfil ao respectivo docente. Como evidências, o sistema utiliza os dados das publicações (o que Ferreira, Gonçalves e Laender (2012) chamam de *citation information*), e todas as informações utilizadas são obtidas da web (dos perfis do Google Acadêmico e dos currículos da Plataforma Lattes).

3 Materiais e métodos

Esta seção apresenta os materiais e métodos utilizados neste trabalho. Foram utilizadas três fontes de dados: uma fonte de dados institucional da Universidade de São Paulo (USP) (UNIVERSIDADE..., 2012a), os currículos da Plataforma Lattes (BRASIL, 1999a) e perfis do Google Acadêmico (GOOGLE, 2004a).

Para esse projeto, foram empregadas informações da USP extraídas do DataUSP, que é um "[...] conjunto de serviços computacionais analíticos para apoio à tomada de decisões da USP." (UNIVERSIDADE..., 2012b, doc. eletrônico). As informações utilizadas nessa base são o nome e o identificador do currículo Lattes dos docentes da USP. Ao todo, foram identificados 8.931 docentes (ativos e inativos) que têm seu respectivo identificador do currículo Lattes cadastrado no sistema[1].

A Plataforma Lattes, mantida pelo CNPq, corresponde à "[...] integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações." (BRASIL, 1999b, doc. eletrônico). Dentre as informações gerenciadas pela Plataforma Lattes estão os currículos Lattes que têm sido usados como um padrão para o registro nacional dos pesquisadores brasileiros e que são amplamente adotados por instituições de pesquisa, universidades e agências de fomento.

Atualmente, a base contém o cadastro de aproximadamente cinco milhões de currículos[2]. Para esse trabalho, foram utilizados os títulos de quatro tipos de produções bibliográficas (livros, capítulos, artigos publicados em periódicos e em

anais de eventos) de cada um dos 8.931 docentes da amostra. Os docentes da amostra apresentavam cadastrados em seus currículos 30.639 livros, 86.793 capítulos de livros, 414.947 artigos em revistas científicas e 578.505 artigos em anais de eventos, totalizando 1.110.884 registros de publicações.

O Google Acadêmico (*Google Scholar* em inglês) é um sistema mantido pela empresa Google que contém um cadastro de informações acadêmicas obtidas da internet (NORUZI, 2007). Além de manter um registro bastante vasto de publicações, o sistema também identifica as citações a cada uma das publicações cadastradas e permite aos autores a criação de perfis acadêmicos que contêm: informações pessoais, como nome do pesquisador, afiliação, áreas de interesse, e-mail e página da web (*home page*) do pesquisador; lista de publicações que pode conter as publicações identificadas automaticamente pelo Google Acadêmico e registros de publicações cadastrados manualmente pelo pesquisador; total de citações de cada uma das publicações; citações recebidas por ano pelo pesquisador, índice h e índice i10 do pesquisador; e lista de coautores do pesquisador.

A maior vantagem do Google Acadêmico para pesquisas bibliométricas ou de análise de grupos acadêmicos é apresentar uma base de dados bastante ampla de publicações (virtualmente, todas as publicações presentes na internet, em sites acadêmicos ou bibliotecas digitais, podem estar indexadas) e de citações (as publicações indexadas são processadas automaticamente por um sistema computacional, de forma a se identificar quais são os trabalhos citados por elas). Adicionalmente, qualquer pesquisador pode criar um perfil no Google Acadêmico, sendo capaz não apenas de completar informações faltantes, corrigir alguns tipos de informações erradas, bem como mesclar registros de publicações (quando uma mesma publicação está armazenada na base, em diferentes registros). O sistema permite que sejam feitas consultas web por publicação e, também, por pesquisador.

Uma das principais desvantagens é a falta de validação dos dados presentes na base do Google Acadêmico, incluindo a falta de validação do resultado dos processamentos automáticos e, também, dos dados cadastrados pelos pesquisadores. Assim, é possível que textos disponíveis na internet sejam

erroneamente registrados como publicações em periódicos ou anais de evento. Além disso, é possível a um pesquisador, em seu perfil, assumir a autoria de publicações que não foram realizadas por ele. Apesar desta desvantagem, as vantagens desse sistema justificam sua utilização, especialmente, se for utilizado como uma fonte de informação complementar para análises bibliométricas, ou análise de grupos de pesquisa.

O método para a identificação do perfil acadêmico de cada pesquisador, proposto e desenvolvido neste projeto, é dividido em quatro partes: busca por pesquisadores no Google Acadêmico, casamento de nomes, casamento de publicações e resolução de nome de autor, propriamente dita.

Busca por pesquisadores no Google Acadêmico: consiste em, dado o nome completo de um docente da amostra, utilizar a ferramenta de busca por perfis, disponibilizada de maneira online pelo sistema (GOOGLE, 2004b), e armazenar as respostas obtidas. Para isto, inicialmente, é feita a busca pelo nome completo do autor entre aspas e removem-se do nome do autor os acentos e os sinais de pontuação como hifens ou apóstrofes. Se esta consulta retornar um ou mais resultados, estes são armazenados e utilizados nas próximas etapas de processamento. Caso nenhum resultado seja retornado, uma nova consulta é executada. Se ainda assim, nenhum resultado for retornado, uma última consulta adicional é executada, e seus resultados são armazenados. A primeira consulta adicional consiste da busca pelo primeiro e último nome do autor (em vez do nome completo), já a segunda, consiste na busca pelo primeiro e penúltimo nome do autor. Apenas para ilustrar, dado um docente hipotético cujo nome é José Maria dos Santos, a busca original seria feita procurando-se "jose maria dos santos". Caso não retornasse nenhum resultado, isto é, nenhum perfil foi encontrado com esse nome, a primeira busca adicional seria feita procurando-se "jose santos" e a segunda procurando-se "jose maria". O quadro 1 resume os resultados do exemplo apresentado para a busca do perfil de José Maria dos Santos. Destaca-se que todas as consultas à ferramenta do Google Acadêmico podem retornar zero, um ou múltiplos perfis. Estes resultados são tratados nas próximas atividades.

Quadro 1 - Exemplo de buscas e resultados no Google Acadêmico

Nome	Busca Realizada	Perfis
"jose maria dos santos"	https://scholar.google.com/citations?view_op=search_authors&mauthors="jose+maria+dos+santos"	0
"jose santos"	https://scholar.google.com/citations?view_op=search_authors&mauthors="jose+santos"	38
"jose maria"	https://scholar.google.com/citations?view_op=search_authors&mauthors="jose+maria"	242

Fonte: Elaborado pelos autores.

Casamento de nomes: consiste na comparação entre o nome completo do docente USP e o nome no perfil do Google Acadêmico encontrado. Essa etapa classifica cada par de nomes comparados como *iguais*, *compatíveis* ou *diferentes*. São considerados *iguais* os nomes que são idênticos após a remoção da acentuação, sinais de pontuação e preposições (por exemplo, de, da, do, dos, das). Para serem considerados *compatíveis*, após a remoção de acentuação e demais filtragens, é necessário que: (i) o nome no perfil acadêmico corresponda a um subconjunto do nome do docente (por exemplo, se no cadastro USP um docente chama-se "José Maria dos Santos", os seguintes nomes filtrados do Google Acadêmico seriam considerados compatíveis por este critério: (1) "jose santos" e "jose maria"; (2) também serão considerados compatíveis os nomes que tenham abreviações cujo nome correspondente não abreviado consta no nome completo do docente. Se no cadastro USP, por exemplo, um docente chama-se "José Maria dos Santos", o seguinte nome filtrado do Google Acadêmico seria considerado compatível por este critério: "jose m santos".

Casamento de publicações: consiste na identificação da existência de títulos de produções bibliográficas de um pesquisador em ambas as bases: Google Acadêmico e Plataforma Lattes. Para cada perfil do Google Acadêmico identificado, na primeira atividade, é feita uma consulta no Google Acadêmico, solicitando-se a lista das 100 primeiras publicações do respectivo perfil, ordenadas da que recebeu mais citações para a que recebeu menos. Assim, para

cada um dos perfis, o sistema terá de zero (caso o pesquisador não tenha nenhuma publicação cadastrada) até 100 publicações. Em seguida, é feita uma busca do título da publicação (após a remoção de acentos e pontuações), nas publicações cadastradas, no currículo Lattes do pesquisador (também, após a remoção de acentos e pontuações). Neste processo é exigido o casamento exato entre o título filtrado do perfil do Google Acadêmico e do currículo Lattes. Para cada docente da USP só é realizado o casamento de publicações entre seu currículo Lattes e os perfis que apresentam nomes *iguais* ou *compatíveis*. Como resultado, para cada um dos potenciais perfis de cada docente, o sistema armazena a quantidade de publicações obtidas do perfil do Google Acadêmico (valor de zero a cem) e a porcentagem dessas publicações que foram encontradas no currículo Lattes do pesquisador (valor de 0% a 100%).

A resolução do nome de autor, propriamente dita, é realizada combinando-se os resultados das atividades anteriores e estabelecendo-se regras para que um perfil seja ou não associado a um dado docente USP. Para cada docente USP, são verificados quantos perfis apresentam nomes iguais ou compatíveis e que apresentam, ao menos, o casamento de uma publicação. Caso haja apenas um perfil que satisfaça tal condição, esse é atribuído ao docente. Caso contrário, atribui-se ao docente o perfil do Google Acadêmico (com nome *igual* ou *compatível*) que apresente a maior quantidade de publicações identificadas no currículo Lattes do pesquisador.

4 Análise dos resultados

Nesta seção são analisados os resultados do método proposto para a resolução de nomes de autores. Todos os resultados obtidos automaticamente pelo método foram validados manualmente.

A partir dos nomes dos 8.931, foram realizadas inicialmente 8.915 consultas diferentes por pesquisadores no Google Acadêmico (a diferença entre o número de docentes e o número de consultas se dá pela existência de homônimos). Destas consultas, 3.316 retornaram um ou mais resultados (37,2%), identificando um total de 4.066 perfis diferentes. Para os docentes cujas consultas não retornaram nenhum perfil, foram realizadas as consultas adicionais, totalizando

4.741 consultas diferentes, nas quais procurou-se apenas pelo primeiro e pelo último nome do docente. Tais consultas identificaram 10.293 perfis. Destaca-se que, após as consultas iniciais, 5.599 docentes ainda não tinham nenhum perfil potencial identificado no Google Scholar, porém, a criação de consultas utilizando apenas o primeiro e o último nome desses docentes deu origem a 4.741 consultas diferentes.

Por fim, 1.965 novas consultas, procurando-se pelo primeiro e pelo penúltimo nome do docente, foram realizadas, e 5.381 perfis foram identificados. Assim, foram realizadas, ao todo, 15.621 consultas por nomes de pesquisadores no Google Acadêmico, sendo identificados, ao todo, 19.740 perfis. Contudo, um mesmo perfil foi identificado por diferentes consultas. No total, 16.135 perfis diferentes foram identificados.

Para cada perfil identificado, foi obtida a lista de suas 100 primeiras publicações (ordenadas da que recebeu o maior número de citações para a que recebeu o menor número, o que corresponde à ordenação padrão fornecida pelo Google Acadêmico). Um total de 832.932 publicações foi obtido, com uma média de 51,6 publicações por perfil.

O processo de casamento entre os nomes de docentes e os nomes nos perfis do Google Acadêmico comparou 27.935 pares de nomes e classificou 3.439 pares como *iguais* e 3.097 como *compatíveis*. Para todos os 6.536 casamentos *iguais* ou *compatíveis,* foi realizado o casamento de publicações. Foram localizadas 298.783 publicações nestes perfis do Google Acadêmico e, destas, 152.303 foram, também, encontradas nos respectivos currículos Lattes.

A resolução de nomes propriamente dita, então, atribuiu a cada docente USP o respectivo perfil no Google Acadêmico, desde que o nome no perfil do Google Acadêmico fosse *igual* ou *compatível,* e houvesse, ao menos, o casamento de uma publicação entre o perfil no Google e o currículo Lattes do docente. Ao todo, 4.613 perfis satisfizeram esse critério, sendo que 3.152 apresentavam os nomes *iguais* aos do docente e 1.461 apresentavam os nomes *compatíveis*. Assim, 4.613 perfis do Google Acadêmico satisfizeram os quesitos de compatibilidade com 4.283 docentes. Para os casos em que mais de um perfil foi considerado *igual*

ou *compatível*, atribuiu-se ao docente o perfil que apresentava o maior número de casamento de publicações.

Ao final do processo, foram encontrados, automaticamente, perfis do Google Acadêmico para 4.283 docentes da USP. Todos estes registros foram validados manualmente e, efetivamente, correspondem aos perfis dos respectivos docentes. Analisando a comparação de nomes, dos 4.283 perfis encontrados, 2.996 apresentam nomes iguais aos nomes dos docentes e 1.287 apresentam nomes compatíveis.

Foram identificadas 278.530 publicações oriundas dos 4.283 perfis e, dessas, 142.429 (51%) casaram com publicações dos currículos Lattes dos respectivos docentes. Essa porcentagem de casamento, relativamente baixa, ocorre por três motivos principais: existem diversas "publicações" nos perfis do Google Acadêmico que não se tratam de capítulos, livros ou artigos científicos e, por isso, não são cadastradas nos currículos Lattes; muitos currículos Lattes não são atualizados com frequência (DIGIAMPIETRI et al., 2014); foi utilizada uma estratégia de casamento simples (casamento exato após a remoção de acentos e pontuações). No entanto, como será discutido adiante, os resultados obtidos foram bastante satisfatórios, não necessitando de etapas mais complexas (e, computacionalmente, mais demoradas) para o casamento das publicações. O quadro 2 sumariza os resultados de identificação de perfis.

Quadro 2 - Sumário dos resultados obtidos

Característica	**Número de perfis**
Docentes na amostra	8.931
Perfis do Google Acadêmico identificados pelas consultas realizadas	16.135
Docentes com perfis no Google Acadêmico, com nomes iguais ou compatíveis	4.283
Perfis do Google Acadêmico com nomes iguais ou compatíveis com o nome dos docentes	4.613
Atribuição de perfis aos docentes após a resolução de nomes	4.283
Número estimado de docentes da USP que possuem perfil no Google Acadêmico (de acordo com a análise de cobertura apresentada a seguir)	4.514

Fonte: Elaborado pelos autores.

O resultado foi também comparado com o cadastro manual de perfis de docentes da USP (que permite a cada docente cadastrar o identificador de seu perfil do Google Acadêmico). O cadastrado manual continha 3.356 perfis do Google Acadêmico.

O sistema automático foi capaz de identificar os mesmos perfis para 3.033 docentes (90,4%); por outro lado, 188 (6,2%) cadastros manuais não foram identificados pela ferramenta automática, porém, 16 deles continham identificadores inválidos de perfis (erros de cadastro), como cadastro do e-mail do docente em vez de seu identificador. Já para os 135 (4,4%) perfis cadastrados manualmente ocorreram divergências entre os resultados obtidos automaticamente e aqueles cadastrados, conforme detalhado a seguir.

A divergência mais comum ocorreu em 102 (3,4%) dos registros, os quais apresentavam identificadores de perfis do Google Acadêmico corretos, porém desatualizados (alguns identificadores dos perfis do Google Acadêmico mudaram ao longo do tempo e assim, dois identificadores diferentes poderiam apontar para o mesmo perfil). Para tais casos, o sistema automático foi capaz de encontrar a versão atual dos identificadores.

Já 26 (0,9%) dos perfis apresentaram erros em seus identificadores (erro de digitação, por exemplo, durante o cadastro manual) e foram identificados, corretamente, pelo sistema automático. Por outro lado, 6 (0,2%) docentes apresentaram dois perfis no Google Acadêmico, e o sistema automático indicou um perfil diferente daquele cadastrado manualmente. Por fim, havia um único registro, cujo cadastro manual apontava para o perfil de outro docente e que o sistema automático foi capaz de identificar o perfil correto. O quadro 3 sumariza essa comparação.

Quadro 3 - Comparação dos resultados automáticos com o cadastro manual

Característica	Número de perfis
Número total de perfis cadastrados manualmente	3.356
Mesmos perfis identificados pelo sistema automático	3.033
Perfis identificados automaticamente com identificadores mais atualizados	102
Perfis manuais cadastrados incorretamente (perfis inválidos)	26
Perfis corretos, mas diferentes (docentes com dois perfis)	6

Cadastro manual apontando para perfil de outro docente	1
Perfis não encontrados pelo sistema automático	188

Fonte: Elaborado pelos autores.

Adicionalmente, o sistema automático foi capaz de identificar o perfil do Google Acadêmico de mais 1.125 docentes que não apresentavam essa informação em seu cadastro.

Dentre 8.931 docentes, o sistema automático foi capaz de identificar, de maneira correta, 4.283 (48%) perfis do Google Acadêmico. O sistema não produziu nenhum falso-positivo, isto é, não atribuiu de maneira incorreta nenhum perfil a um dado docente (com a única ressalva que seis docentes apresentavam dois perfis no Google Acadêmico, e o sistema identificou um perfil diferente daquele cadastrado manualmente pelo pesquisador). Dos 3.356 perfis que haviam sido cadastros manualmente, o sistema não foi capaz de encontrar apenas 188 (6,2%), sendo que 16 desses cadastrados apresentaram erros. Assim, apenas 172 (pouco mais de 5%) dos perfis corretamente cadastrados, de maneira manual, não foram identificados automaticamente. Uma verificação manual dos resultados permitiu identificar que o principal motivo para estes registros não terem sido identificados automaticamente ocorreu no processo de casamento de nomes. Tal processo considerou os nomes desses perfis incompatíveis com os nomes dos docentes, por possuírem, por exemplo, apelidos ou siglas de instituições juntamente com o nome do pesquisador.

Destaca-se que nem todos os docentes apresentam um perfil no Google Acadêmico e, pela validação realizada, observa-se que a ferramenta foi capaz de encontrar cerca de 95% dos perfis daqueles docentes que, pelo cadastro manual, sabidamente têm perfis acadêmicos.

6 Considerações finais

Este artigo apresentou uma estratégia que combina dados de três fontes diferentes para a resolução do nome de autores. Especificamente, foi tratado o problema de identificar, de maneira automática, o perfil do Google Acadêmico de docentes com base em seu nome completo e sua lista de publicações extraída da plataforma Lattes.

A estratégia é baseada na busca pelos perfis dos docentes no Google Acadêmico, considerando o nome completo do docente, o primeiro e o último nome e o primeiro e o penúltimo nome. Adicionalmente, é realizado um processo de casamento de nomes, a fim de verificar a compatibilidade entre o nome buscado e o nome encontrado, e são comparadas as publicações acadêmicas que estão cadastradas no currículo Lattes do docente e no perfil do Google Acadêmico. Por fim, a resolução de nomes ocorre verificando-se entre os perfis compatíveis aquele que apresenta maiores evidências de pertencer ao respectivo docente.

A validação realizada nos 4.283 perfis encontrados pela ferramenta mostrou que não ocorreu nenhum falso-positivo, ou seja, cada um destes perfis pertencia ao respectivo docente. A única ressalva é que, para seis docentes, o sistema automático identificou corretamente o respectivo perfil do Google Acadêmico, porém, estes docentes apresentam mais de um perfil e, no cadastro manual, haviam cadastrado um perfil diferente daquele encontrado pelo sistema.

Além disso, uma análise de cobertura mostrou que, considerando a base de registros de perfis cadastrados manualmente, o sistema foi capaz de identificar cerca de 95% deles.

Como trabalhos futuros, pretende-se aumentar a cobertura do sistema, de forma a tratar os desafios de atribuir corretamente um perfil a um docente, considerando que o nome cadastrado no perfil pode ser diferente do nome do docente (possuindo, por exemplo, um apelido ou mesmo o nome ou a sigla da instituição ou departamento no qual o pesquisador trabalha).

Agradecimentos

Agradecemos à Universidade de São Paulo por prover a infraestrutura necessária para a realização da pesquisa apresentada neste artigo e à equipe da Superintendência de Tecnologia da Informação (STI) da USP.

Referências

BORGES, Eduardo et al. An unsupervised heuristic-based approach for bibliographic metadata deduplication. **Information Processing & Management**, New York, v. 47, n. 5, p. 706-718, Sept. 2011.

BRASIL. Conselho Nacional de Desenvolvimento Científico e Tecnológico. **Plataforma Lattes**. 1999a.

BRASIL. Conselho Nacional de Desenvolvimento Científico e Tecnológico. **Plataforma Lattes**: sobre a plataforma. 1999b.

CANUTO, Sérgio et al. UDRB: Uma nova heurística eficaz para deduplicação de referências bibliográficas. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 28., 2013, Recife. **Anais...** Recife: UFPE, 2013. p. 1-6.

DIAZ-VALENZUELA, Irene; MARTÍN-BAUTISTA, Maria J.; VILA, Maria A. A fuzzy semisupervised clustering method: Application to the classification of scientific publications. In: INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING AND MANAGEMENT OF UNCERTAINTY IN KNOWLEDGE-BASED SYSTEMS, 15., 2014, Montpellier. **Anais...** Montpellier: IPMU, 2014. p. 179-188.

DIGIAMPIETRI, Luciano A. et al. Análise macro das últimas atualizações dos currículos Lattes. **Em Questão**, Porto Alegre, v. 20, n. 3, p. 88-113, 2014. Edição Especial.

DIGIAMPIETRI, Luciano A.; BARBOSA, Lênin F.; LINDEN, Ricardo. Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 4., 2015, Porto Alegre. **Anais...** Porto Alegre: PUCRS, 2015. [p. 1-6].

FERREIRA, Anderson A.; GONÇALVES, Marcos A.; LAENDER, Alberto H. F. A brief survey of automatic methods for author name disambiguation. **Sigmod Record**, New York, v. 41, n. 2, p. 15-26, June 2012.

FERREIRA, Anderson A.; MACHADO, Tales M.; GONÇALVES, Marcos. A. Improving author name disambiguation with user relevance feedback. **Journal of Information and Data Management**, [S.l.], v. 3, n. 3, p. 332-347, Oct. 2012.

GODOI, Thiago A. et al. A relevance feedback approach for the author name disambiguation problem. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, 13., 2013, New York. **Proceedings…** New York: ACM/IEEE-CS, 2013. p. 209-218.

GOMIDE, Janaina; KLING, Hugo; FIGUEIREDO, Daniel. Name usage pattern in the synonym ambiguity problem in bibliographic data. **Scientometrics**, Dordrecht, v. 112, n. 2, p. 747-766, Aug. 2017.

GOOGLE. **Google Acadêmico.** 2004a.

GOOGLE. **Google Acadêmico**: busca por autores. 2004b.

HAN, Hui et al. Two supervised learning approaches for name disambiguation in author citations. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, 4., 2004, Tucson. **Proceedings…** Tucson: ACM/IEEE-CS, 2004. p. 296-305.

MILOJEVIC, Stasa. Accuracy of simple, initials-based methods for author name disambiguation. **Journal of Informetrics**, Amsterdam, v. 7, n. 2, p.767-773, 2013.

MUGNAINI, Rogério et al. Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros. **Em Questão**, Porto Alegre, v. 18, n. 3, p. 263-279, dez. 2012. Edição especial.

NORUZI, Alireza. Google scholar: The new generation of citation indexes. **Libri**, Berlin, v. 55, n. 4, p. 170-180, 2007.

NOVÁK, Vilém; PERFILIEVA, Irina; MOCKOR, Jiri. **Mathematical principles of fuzzy logic**. [S.l.]: Springer Science & Business Media, 2012.

SONG, Wanpeng et al. Question similarity calculation for FAQ answering. In: INTERNATIONAL CONFERENCE ON SEMANTICS, KNOWLEDGE AND GRID, 3., 2007, Washington. **Anais…** Washington: IEEE Computer Society, 2007. p. 298-301,

STROTMANN, Andreas; ZHAO, Dangzhi. Author name disambiguation: What difference does it make in author-based citation analysis? **Journal of the American Society for Information Science and Technology**, Hoboken, v. 63, n. 9, p. 1820-1833, Aug. 2012.

TALBURT, John R. **Entity resolution and information quality**. San Francisco: Morgan Kaufmann, 2010.

UNIVERSIDADE DE SÃO PAULO. Superintendência de Tecnologia da Informação. **DataUSP**. 2012a.

UNIVERSIDADE DE SÃO PAULO. Superintendência de Tecnologia da Informação. **DataUSP**: apresentação. 2012b.

WHITE, Howard D.; MCCAIN, Katherine W. Visualizing a discipline: An author co-citation analysis of information science 1972-1995. **Journal of the American Society for Information Science**, New York, v. 49, n. 4, p. 327-355, Dec. 1998.

Automatic identification of academic profiles using author name disambiguation

Abstract: The author name disambiguation is a fundamental activity in bibliometric studies, in particular in those that use different sources of information. The objective of this paper is to propose and test an author name disambiguation strategy in order to allow the automatic identification of the Google Academic profile of researchers. The proposed strategy is based on the search for the profiles in Google Scholar, followed by a name matching process. Additionally, the academic publications that are registered in the researcher's Lattes curriculum and Google Scholar profile are compared. Lastly, the name resolution is carried out by verifying among the compatible profiles the one with the highest evidence of belonging to the respective researcher. A case study involving researchers from the University of São Paulo was conducted, and the automated system was able to correctly identify 4,283 Google Scholar profiles. A coverage analysis showed that the system was able to find about 95% of the profiles of the researchers who have this information, and no false-positive was identified.

Keywords: Author name disambiguation. Entity resolution. Bibliometrics.

Recebido: 09/06/2017

Aceito: 20/08/2017

[1] A USP tem 5.903 docentes ativos, sendo que destes, 5.876 possuem o identificador de seus currículos cadastrado (UNIVERSIDADE, 2012a).

[2] No dia 30 de janeiro de 2017, existiam 4.996.326 currículos cadastrados na Plataforma Lattes (BRASIL, 1999a).