• IJKEM, INT. J. KNOWL. ENG. MANAGE., v.5, n.13 • FLORIANÓPOLIS, SC • NOV. 2016/FEV. 2017• ISSN 2316-6517 •
Submissão: 20 set. 2016. **Aceitação**: 17 out. 2016. **Sistema de avaliação**: às cegas dupla (*double blind review*).
UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC)
João Artur de Souza e Gertrudes Aparecida Dandolini (Ed.), p. 42-57.

ALINHAMENTO DE NOMES DE COAUTORES DE PUBLICAÇÕES CIENTÍFICAS: UMA ABORDAGEM PRÁTICA

ANDRÉ GUIMARÃES PEIL
Bacharel em Ciência da Computação, Universidade Federal de Pelotas,
agpeil@inf.ufpel.edu.br

DANIELA FRANCISCO BRAUNER
Doutora em Informática pela Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio,
Professora adjunta, Escola de Administração, Universidade Federal do Rio Grande do Sul.
daniela.brauner@ufrgs.br

RICARDO MATSUMURA DE ARAÚJO
Doutor em Ciência da Computação pela UFRGS.
Professor adjunto, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.
ricardo@inf.ufpel.edu.br

GLAUCO ROBERTO MUNSBERG DOS SANTOS
Mestrando em Ciência da Computação, Universidade Federal de Pelotas,
grmdsantos@inf.ufpel.edu.br

RESUMO

Objetivo: Este trabalho propõe um processo para o alinhamento de nomes para apoiar o tratamento da ambiguidade de coautorias em produções científicas, tais como publicações.

Design/Metodologia/Abordagem: O artigo apresenta uma abordagem prática, aplicada a uma demanda para suporte às avaliações anuais dos programas de pós-graduação brasileiros, usando dados oriundos da Plataforma Lattes.

Resultados: Os resultados obtidos no contexto do experimento foram considerados satisfatórios, foi possível encontrar dois valores nos quais permitem ter um grande número de respostas positivas. No entanto, devido a baixa quantidade de informações que temos com relação aos coautores nos currículos lattes, não há como garantir uma exatidão nas descobertas dos alinhamentos.

Originalidade/valor: Esta abordagem é parte de um estudo que visa desenvolver um sistema de análise gráfica da produção científica brasileira, com o propósito de fornecer a coordenadores e gestores de um curso ou universidade, uma ferramenta na qual seja possível fazer um acompanhamento das produções científicas de forma dinâmica.

Palavras-chave: Alinhamento. Desambiguação. Web-semântica. Distância de edição.

ALIGNING CO-AUTHORS NAMES IN SCIENTIFIC PUBLICATIONS: A PRACTICAL APPROACH

ABSTRACT

Purpose: This paper proposes a method for alignment of names to support the treatment of co-authorships ambiguity in scientific productions such as publications.

Design/Methodology/Approach: The paper presents a practical approach, applied to a demand for support for annual assessments of the Brazilian graduate program, using data from the Lattes Platform.

Results: The results obtained in the experimental context were considered satisfactory, it was possible to find two values which allow to have a large number of positive responses. However, due to the low amount of information we have regarding the co-authors in lattes curriculum, there is no guarantee accuracy in the alignments findings.

Originality/value: This approach is part of a study to develop a graphic analysis system of the scientific production, with the purpose of providing the engineers and managers of a course or university, a tool in which to make a follow-up productions scientific dynamically.

Keywords: Alignment. Disambiguation. Web-semantic. Distance edition.

1 INTRODUÇÃO

Os programas de pós-graduação brasileiros necessitam periodicamente analisar o rendimento da produção acadêmica dos pesquisadores, grupos de pesquisa, projetos entre outros. Estes balanços são feitos anualmente com base em métricas e quantificações definidas pela CAPES. Os dados são levantados dos currículos Lattes e os cálculos realizados com o auxílio de sistemas externos. Tais sistemas recebem como entrada currículos, oriundos da Plataforma Lattes mantida pelo CNPq, com dados dos pesquisadores participantes dos programas. Internamente é importante que estes softwares de análise de produção bibliográfica, se preocupem com problemas de ambiguidades de publicações, de nomes de autores, nomes de eventos, entre outros. O somatório das produções de um grupo de pesquisadores que possuem publicações em comum deve ser feito eliminando-se manualmente as duplicatas.

O objetivo deste trabalho é fazer alinhamentos entre nomes de autores em grupos de currículos pré-selecionados, usando como premissas suas publicações em comum a fim de auxiliar a deduplicação de publicações de forma automatizada. A deduplicação é o processo de analisar os dados e eliminar redundâncias, diminuindo assim a quantidade de informação a ser manipulada. Ela é importante, neste cenário, para aplicações que computam métricas de desempenho dos programas de pós-graduação. A identificação de autores nas coautorias de publicações científicas é igualmente importante para a criação de redes de relacionamento. Como resultado da abordagem proposta, é apresentado um valor de similaridade entre os autores. Um pesquisador citado como *João Silva da Silva* comparado com outra instância representada por *João S. Silva*, receberá um valor de similaridade, que poderá sugerir que ambas instâncias se referem a mesma pessoa. Este artigo apresenta também um estudo de caso prático de aplicação desta proposta, bem como a validação dos valores encontrados. São utilizados currículos de pesquisadores da Universidade Federal de Pelotas e os resultados foram implementados numa plataforma de análise gráfica e numérica dos currículos de pesquisadores chamada Cientum[1].

Este artigo é organizado da seguinte forma. Desenvolvimento em um primeiro momento, discutimos trabalhos relacionados e um segundo momento detalhamos a nossa abordagem ao problema, posteriormente apresentamos experimentos realizados para validação da abordagem. E por fim, é apontando as principais contribuições e possibilidades de trabalhos futuros.

[1] http://cientum.indeorum.com

2 DESENVOLVIMENTO

O problema de ambiguidade de nomes tem sido estudado em vários contextos, como desambiguação de termos biomédicos Al-Mubaid e Chen, (2006), nomes geográficos Casanova, (2008) e Brauner, (2007), no contexto de desambiguação em bases bibliográficas na Web temos muitos estudos utilizando a *Digital Bibliography and Library Project* (DBLP)[2] , MEDLINE[3], Biblioteca Digital Brasileira de Computação (BDBC)[4] e CiteSeer[5]. Segundo (Felipe Levin, 2010) o problema vem recebendo diferentes denominações ao longo dos anos: *merge/purge*, Hernández e Stolfo, (1995), *Record linkage* Chaudhuri, et al. (2003), *duplicate detection* Weis e Naumann, (2006) e *reference desambiguation* Kalashnikov e Mehrotra, (2006).

As bibliotecas digitais de publicações científicas, por possuírem conjuntos de publicações de diversos grupos e áreas, apresentam a problemática de ambiguidade de nomes de coautores bastante evidente, como por exemplo: "SILVA, A.", que tanto se refere a "André Silva" quanto a "Alexandre Silva". Assim como nas bibliotecas digitais, quando se extrai um grupo de currículos da Plataforma Lattes para fazer uma análise, acaba-se tendo o mesmo problema. Porém Bibliotecas Digitais são diferentes de Currículos Lattes. No entanto a desambiguação no cenário de Currículos Lattes é essencial para, por exemplo, o cálculo da produtividade dos programas de pós-graduação.

As abordagens não supervisionadas não usam exemplos de conjuntos de treinamento e um classificador como as supervisionadas, mas exploram atributos das publicações para criar agrupamentos baseados em similaridades de tais atributos, indicando mesma autoria Sun et al. (2011). O HHC*,* definido como *Heuristical Hierarchical Clustering,* Cota et al, (2007) é um método baseado nos atributos de citações presentes em bibliotecas digitais, tais como título da publicação, local da conferência, ano e lista dos coautores. O objetivo do HHC é agrupar citações em conjuntos que correspondem ao mesmo autor. Para isso, ele utiliza heurísticas como funções de similaridade. De forma geral, o algoritmo utiliza duas *strings*, correspondendo a dois nomes de autores, e as compara por meio de uma função de distância de edição. Se os dois nomes são considerados equivalentes para um determinado valor limite, então os demais dados, como nomes de coautores, títulos e locais de publicação são usados como evidências adicionais para determinar se dois nomes correspondem a um mesmo autor Cota et al, (2007). Neste artigo, propomos uma abordagem similar, pois também agrupamos por similaridade, porém a heurística

[2] http://dblp.uni-trier.de
[3] http://www.ncbi.nlm.nih.gov/pubmed/
[4] http://www.lbd.dcc.ufmg.br/bdbcomp/
[5] http://citeseer.ist.psu.edu/index

aplicada para os agrupamentos é feita em vários níveis e não diretamente em cima de autores similares.

O *Author-ity*, é um modelo para estimar a probabilidade de dois artigos que compartilham os mesmos autores Torvik e Smalheiser (2009). O estudo de caso foi feito na base bibliográfica da MEDLINE propõe o agrupamento através de um algoritmo guloso que começa com todos artigos da base e vai mesclando-os de forma interativa, através do cálculo de similaridade sintática dos nomes dos autores, até que não haja mais possibilidade de agrupá-los. Este trabalho inspirou nossa proposta pois divide o processo em uma metodologia baseada em etapas, mas difere na abordagem de agrupamento, pois ao preferimos ir reduzindo o escopo de comparações através de agrupamentos intermediários.

Com relação as atuais ferramentas que lidam com dados extraídos da Plataforma Lattes, é importante citar o *scriptLattes* Mena-Chalco et al. (2009). O *scriptLattes* é uma ferramenta de código fonte aberto que extrai dados de um conjunto de currículos Lattes e gera relatórios a partir deles. O sistema utiliza como base do processo de eliminação de duplicatas a métrica de distância de edição (Levenshtein, 1966). O algoritmo determina a similaridade entre duas cadeias de caracteres a partir do número mínimo de edições (inserção, exclusão, substituição) necessárias para transformar uma *string* A em uma *string* B. Outra abordagem similar são os *n-grams*, apresentada em Elmagarmid et al. (2007). Um *n-gram* é um conjunto de **n** caracteres adjacentes. Para calcular a similaridade de *n-grams* entre duas *strings* é necessário contar o número de *n-grams* que aparecem em ambos os *strings*, dividindo este total pelo número total de *n-grams* distintos. Por ser mais preciso, inclusive na falta de padronização na escrita dos nomes de citação, que acontece nos dados utilizados na abordagem proposta neste artigo, foi escolhido o algoritmo de *trigram* (usando cada três caracteres adjacentes) para calcular a similaridade entre o nome dos autores. O algoritmo de Levenshtein é utilizado para comparação entre títulos das publicações.

2.1 ABORDAGEM PROPOSTA

O processo de alinhamento proposto é organizado em duas etapas que serão apresentadas nesta seção. A primeira etapa, chamada Padronização, consiste em coletar transformar e inserir os dados na base. Na segunda etapa, chamada Desambiguação, são realizadas as etapas do processo de alinhamento de fato. Os dados utilizados nos exemplos, são currículos extraídos da Plataforma Lattes[6].

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

2.1.1 Etapa I: Coleta e Padronização

A primeira etapa consiste na coleta do conjunto de dados. Uma das heurísticas do processo proposto se baseia no uso de grupos de pesquisadores que devem possuir relações entre si, de forma a facilitar a desambiguação por apresentarem publicações em comum, i.e., referências bibliográficas em comum em seus currículos. Para isso, é selecionado um conjunto de dados de currículos que possuem alguma relação. Por exemplo, grupos de pesquisadores de uma mesma área científica, de mesmo grupo de pesquisa ou um mesmo programa de pós-graduação. Isto garante, que existirão relações entre algumas publicações deste grupo pré-selecionado. O conjunto de dados utilizado neste artigo foi criado com base em currículos de pesquisadores de uma mesma universidade, com um conjunto limitado de currículos de três laboratórios distintos.

A etapa de padronização é onde foram realizadas atividades de transformação dos dados para um formato que facilite a manipulação das informações. Com base na amostra de dados selecionada, foram realizadas análises na estrutura interna dos Currículos Lattes, com o objetivo de conhecer como são organizados e identificar problemas oriundos de erros de preenchimento dos usuários ou até mesmo falhas/inconsistências da plataforma. Foi utilizado um conjunto de dados de três grupos de pesquisa da Universidade Federal de Pelotas (UFPel), contendo ao todo 116 currículos. Dentre os problemas identificados estão: erros de digitação, falta de informações sobre publicações e erros na conexão entre coautores. Neste último, muitos links foram criados para pessoas que não são os verdadeiros coautores, muito provavelmente devido a seleção manual oferecida na Plataforma Lattes ou por problemas de transformação em dados legados.

Como um dos objetivos do trabalho é retroalimentar os currículos com informações semânticas sobre coautorias para alimentar a ferramenta Cientum, escolhemos transformar previamente os dados da Plataforma Lattes para dados que permitissem a fácil manipulação e a representação semântica das relações de similaridade encontradas para alinhamento de coautorias. Neste sentido, optamos por transformar os dados para a linguagem RDF[7]. Para isto, foi escolhida a ontologia definida no projeto da rede VIVO Rademaker e Haeusler, (2013) como modelo de dados para organização da informação. A VIVO é uma rede interdisciplinar desenvolvida em 2003 e implementada na Cornell University, cujo objetivo é apoiar a pesquisa, a navegação e a visualização de currículos de pesquisadores, a fim de permitir a descoberta de pessoas, programas, instalações, financiamento, trabalhos acadêmicos e eventos.

[7] A Resource Description Framework – RDF, (W3C, 2003)

A Plataforma Lattes exporta seus currículos em formato XML. Neste contexto, foi necessário o uso de um script para converter os dados dos Currículos Lattes de XML[8] para o formato RDF. Para isso foi escolhido o *script Semantic Lattes* ou *Slattes* Rademaker e Haeusler, (2013). Este *script* transforma os dados do Lattes em RDF, seguindo o padrão utilizado pela ontologia VIVO. Todo o processo de alinhamento foi desenvolvido com base na ontologia VIVO e pode ser aproveitado por outros usuários deste padrão de representação de dados.

2.1.2 Etapa 2: Desambiguação

O processo de desambiguação é feito através de três etapas: Normalização, Clusterização e Pareamento Peil et al.,(2015).

Inicialmente, na etapa de normalização dos dados, os dados são padronizados de forma a facilitar as comparações sintáticas. Nesta etapa é realizada uma varredura sintática, para remoção de *stopwords*[9] e diferenças sintáticas. Após isto é feita a clusterização, que realiza o agrupamento de conjuntos de coautores. Inicialmente, os agrupamentos são realizados de acordo com os artigos em comum. Artigos com títulos similares presentes no conjunto de dados representam que os diferentes titulares dos currículos (do conjunto de dados) apresentam relações de coautorias, bastando identificá-los devidamente. Por último, é realizado o pareamento, onde pares de coautores são identificados e valores de similaridade são atribuídos. As próximas seções explicam em detalhes as etapas do processo de desambiguação.

2.1.2.1 NORMALIZAÇÃO

A primeira etapa do processo de desambiguação caracteriza-se como uma limpeza dos dados ou pré-processamento. Esta etapa garante a minimização de erros como troca de ‘ç’ por ‘c’, ‘ã’ por ‘a’ e a substituição de caixa alta (maiúsculas) por caixa baixa (minúsculas). Além disso são retiradas *stopwords*, espaços duplicados e conectivos. Na tabela 1 são apresentados exemplos pertencentes às amostras de dados utilizadas neste trabalho, com o objetivo de exemplificar a etapa de normalização dos dados.

[8] eXtensible Markup Language - http://www.w3schools.com/xml/
[9] Palavras irrelevantes, Álvarez (2007).

Quadro 1 – Exemplo do processo de normalização com entradas e saídas

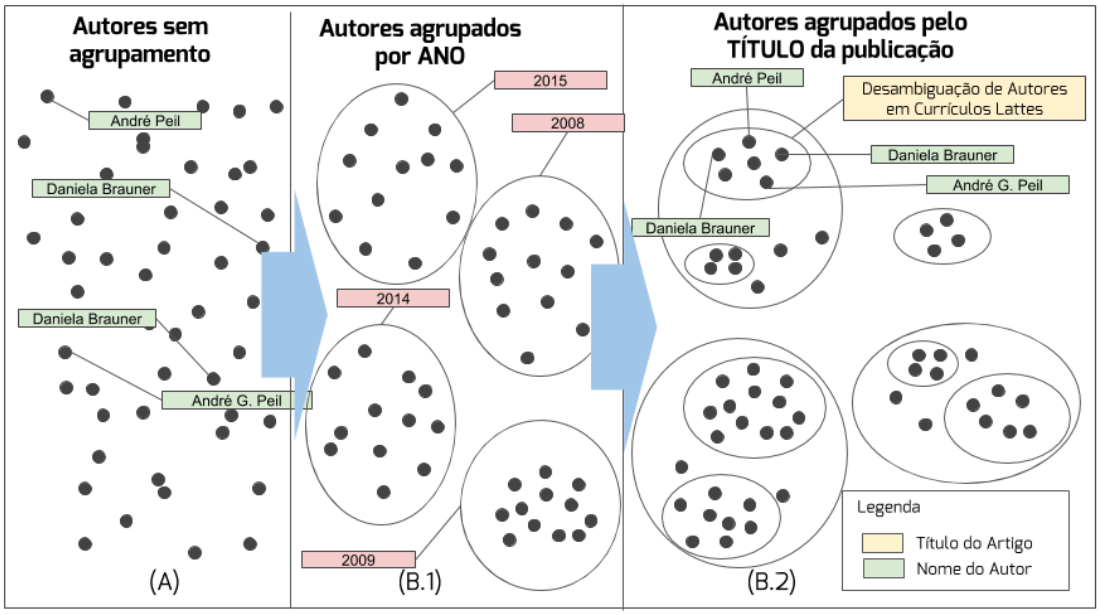
ENTRADA	⇒	SAÍDA
Habitação de interesse social: conceito, método e técnicas.	⇒	habitacao-interesse-social-conceito-metodo-tecnicas
Nova Extensão para a Ferramenta FlexMap Capaz de Explorar	⇒	nova-extensao-para-ferramenta-flexmap-capaz-explorar
Julio Saraçol Domingues Júnior	⇒	julio-saracol-domingues-junior
59º Congresso Brasileiro de Genética	⇒	59-congresso-brasileiro-genetica
Plataforma PEOPLEGRID incluindo a opinião das pessoas no planejamento urbano	⇒	plataforma-peoplegrid-incluindo-opiniao-pessoas-planejamento-urbano

Fonte: Peil et al. (2015).

2.1.2.2 CLUSTERIZAÇÃO

A clusterização ou agrupamento consiste em criar grupos de entidades que se correlacionam seguindo algum critério. Inicialmente foram criados clusters de artigos com mesmo ano de publicação e depois eles foram subdivididos em clusters de artigos com similaridade de título da publicação. A Figura 1 ilustra os agrupamentos realizados. A similaridade dos títulos é realizada com apoio do algoritmo de distância de edição Levenshtein (1966), o qual retorna um valor entre 0 e 1 indicando a proximidade entre os elementos comparados, chamado neste artigo de **α*Título***, explicado na seção 4.

Figura 1 – Representação da formação dos grupos com base na clusterização.



Fonte: Peil et al. (2015).

2.1.2.3 PAREAMENTO

Com base nos clusters de artigos similares, (i.e. cada cluster contém um conjunto de artigos que são similares sintaticamente), teremos um conjunto de autores candidatos aos pareamentos. O pareamento é a etapa final do processo de desambiguação. O objetivo desta etapa é comparar

os autores de cada cluster e atribuir valores de similaridade, calculado a partir de indícios de similaridade. Um primeiro indício de similaridade é a comparação sintática entre o nome dos autores. Ela é feita através do algoritmo de *n-grams* Elmagarmid et al. (2007), que assim como o algoritmo de Leveshtein retorna um valor entre 0 e 1 conforme a similaridade dos nomes comparados, este índice foi denominado **αNome**. Em seguida, o processo calcula um valor de similaridade para cada dupla de autores através da análise de outros atributos. O cálculo do valor de similaridade é feito com base nas informações extraída de um determinado currículo, conforme a Tabela 2. A coluna **Peso** apresenta os pesos atribuídos para os atributos de cada autor. O valor de similaridade (**VS**) é calculado usando a seguinte fórmula: *VS=(A*B)+C+D-(D*E))+F*, cujos operadores são apresentados na Tabela 3. Como resultado, os valores atribuídos para cada par candidato gerará um valor de 0 a 10. A similaridade sintática é utilizada para os itens de título de artigo e nome do autor, depois disso, é verificado, se as publicações respeitam a mesma ordem de autoria e o ano de publicação. Conforme a maior quantidade de atributos iguais entre os autores comparados, maior o valor de similaridade, demonstrando um provável alinhamento entre tais autores.

Quadro 2 – Relação de elementos extraídos dos Currículos Lattes e pesos atribuídos pelo pareamento

CAMPO	DESCRIÇÃO	PESO
refBy	É o identificador que representa o currículo onde foi citado aquele autor.	-
refArticle	É o identificador que representa o artigo de um currículo. Um currículo possui vários artigos logo é necessário identifica-los para não haver conflito entre os autores.	-
refAuthor	É uma URL de identificação do autor	-
nameAuthor	Nome do autor	6
rank	Cada publicação possui uma ordem de autoria entre seus autores	2
nameArticle	Titulo do artigo	1
nameConference	Nome da conferência onde foi publicado	-
year	Ano da publicação	1

Fonte: Peil et al. (2015).

Quadro 3 – Descrição de cada elemento da fórmula

CAMPO	DESCRIÇÃO
A	É o identificador que representa o currículo onde foi citado aquele autor.
B	É o identificador que representa o artigo de um currículo. Um currículo possui vários artigos logo é necessário identifica-los para não haver conflito entre os autores.
C	É uma URL de identificação do autor
D	Nome do autor
E	Cada publicação possui uma ordem de autoria entre seus autores
F	Título do artigo

Fonte: Peil et al. (2015).

3 VALIDAÇÃO E RESULTADOS

A validação deste trabalho é uma etapa importante para verificar se os valores de pesos e os limiares atribuídos para sugestão de similaridades sintáticas utilizadas estão obtendo bons resultados no pareamento de coautorias. O processo de validação inicia passa pela rotulagem dos dados, onde foram marcados todos os casamentos entre autores possíveis dentro de um grupo de currículos, identificando se cada par é ou não um casamento válido. Este processo foi realizado de forma manual, onde todas as possibilidades de casamento foram analisadas e rotuladas como verdadeira (1) e falso (0), em cima de conjuntos de dados apresentados na Tabela 4.

A calibragem do **αTítulo** foi realizada em cima do conjunto B e a do **αNome** com o conjunto A. Num terceiro momento foram testados os limiares obtidos no grupo de pesquisa C, apresentados ao final desta seção.

Quadro 4 - Informações dos Grupos de Pesquisa utilizados

GRU-PO	TOTAL DE CURRÍCULOS DE PESQUISADORES	TOTAL DE PUBLICAÇÕES	TOTAL DE AUTORES AMBÍGUOS	MEDIA DE AUTOR POR PUBLICAÇÃO	QUANTIDADE DE PARES ROTULADOS
A	48	2353	12323	5	20503
B	18	539	1738	3	4198
C	50	3401	14290	4	44562

Fonte: Peil et al. (2015).

Na sequência, foram identificadas situações de similaridade que irão apoiar o processo de validação. A Tabela 5 mostra diferentes situações onde o processo a ser validado irá considerar os acertos (verdadeiro) e os erros (falsos) no pareamento dos coautores. As situações de similaridade definidas têm o objetivo de identificar a ocorrência de falsos positivos e negativos durante a validação, pois existem situações em que a similaridade entre os pares sugere que eles são corretos porém fazendo a combinação entre os atributos verifica-se que são pessoas distintas.

Erros de preenchimento na citação de coautorias podem gerar falsos negativos aplicando-se nossa abordagem. Por exemplo uma citação ao coautor "J. C. Guimarães", que também aparece como "Joaquim da Cunha Guimarães", no conjunto rotulado este par constaria como um casamento verdadeiro, ou seja, representam a mesma pessoa. Entretanto, de forma automatizada, este tipo de comparação de *string*, usando *trigram*, por exemplo, daria um valor alto, fazendo com que fosse considerado como falso, e, portanto, um falso negativo, enquadrado na situação 4. Para termos falsos positivos teríamos que ter autores diferentes, com nomes estritamente parecidos dentro da mesma publicação, de um mesmo grupo (Situação 3). É interessante destacar que esta situação é frequente quando há uma grande incidência de pesquisadores japoneses com nomes

parecidos Torvik e Smalheiser, (2009) ou a ocorrência de irmãos com nomes de citação parecidos em uma mesma publicação, por exemplo Bruno Costa Moura e Bruna Costa Moura.

Quadro 5 – Situações e suas condições

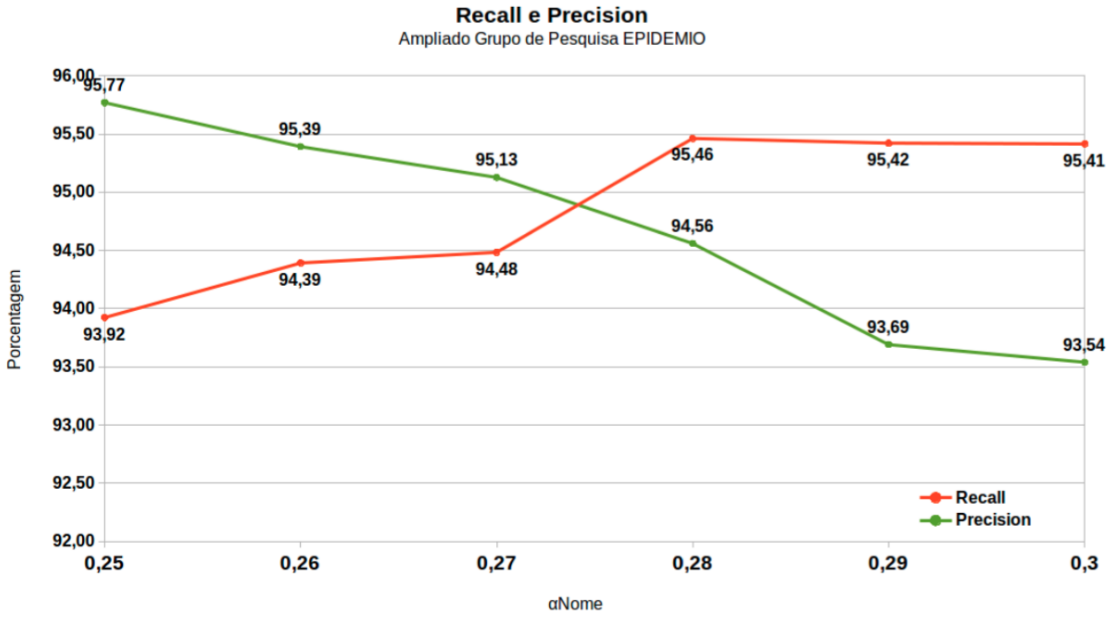
SITUAÇÃO	PAREAMENTO	DESCRIÇÃO
Situação 1	Verdadeiro	Autores com nomes iguais. (A = B)
Situação 2	Verdadeiro	Autores que possuem valor de trigram >= αNome e ordem de autoria correta
Situação 3	Verdadeiro	Autores que possuem valor de trigram >= αNome e ordem de autoria incorreta
Situação 4	Falso	Autores que possuem valor de trigram < αNome e ordem de autoria correta
Situação 5	Falso	Autores que possuem valor de trigram < αNome e ordem de autoria incorreta

Fonte: Peil et al. (2015).

Uma etapa de calibragem foi realizada utilizando-se um conjunto de testes diferente do conjunto de validação para descobrir quais limiares dos algoritmos utilizados para cálculo de similaridade sintática são mais adequados para aplicar a validação dos resultados. Neste caso, a calibragem foi realizada para os dois índices utilizados no processo de desambiguação. Para o índice **αTítulo**, utilizado pela análise de similaridade sintática dos títulos de publicações na etapa de clusterização, foi realizada uma verificação manual dos valores atribuídos aos pares de títulos e foi possível perceber que com o limiar até 0,04 não houve sobreposição de artigos diferentes. Portanto, o valor de **αTítulo** (Leveshtein) foi fixado em 0,04 para validar as outras variáveis do processo.

O índice **αNome**, é atribuído à comparação entre os pares de nomes dos autores durante a etapa de pareamento. Para validar este índice utilizamos outro conjunto de testes, diferente do conjunto utilizado na calibragem anterior. A Figura 2 apresenta o gráfico de *Recall e Precision* do **αNome**. Analisando o gráfico percebemos que o nível de precisão de acertos diminui conforme aumenta o valor de **αNome**. Mas a sensitividade aumenta, conforme evidenciado pelo *Recall*. Portanto, deixando o nível muito baixo do *recall*, aumentamos a quantidade de erros dentro do conjunto de acertos e por isso escolhemos o cruzamento do *recall e precision* como melhor índice para o **αNome**, que ficou entre 0,27 e 0,28, e definimos o melhor em termos de precisão, ficando o **αNome** calibrado em 0,28.

Figura 2 – Título.



Fonte: Peil (2015).

Aplicando-se os limiares definidos sobre o conjunto de dados C, a abordagem proposta obteve 100% de acertos. O resultado nos intrigou e analisando as situações (vide Quadro 6), vimos que o resultado se deve à baixa ocorrência de ambiguidades de nomes nas publicações do grupo e ao baixo erro nas citações entre os currículos dos autores, ocorrendo a maioria dos casos na situação 1.

Quadro 6 – Percentual de acerto de cada situação com o αNome em 0,28.

SITUAÇÃO	TOTAL DE ACERTOS	ACERTOU	ERROU	PORCENTA-GEM
Situação 1 (IGUAIS)	8030	4548	0	56,64%
Situação 2 (αNome e Ranking)	8030	3108	5	38,70%
Situação 3 (αNome)	8030	374	18	4,66%

Fonte: Peil et al. (2015).

Por este motivo, resolvemos aplicar a abordagem em cima de todos os conjuntos de dados (grupos: A,B e C). Os resultados obtidos são apresentados a seguir com base na avaliação das situações para os diferentes grupos. Aparentemente a média do Valor de similaridade (Média) obtido não difere muito em relação a situação como pode ser visto nas Tabelas 7 a 11.

O que de fato influencia se o casamento entre os autores é verdadeiro ou não é o valor de similaridade final estar acima da média das situações que são consideradas verdadeiras. O fato das médias da situação 3 estar entre *5,2 e 5,4*, nos leva a concluir que pareamentos com estes valores tem grande chance de serem verdadeiros para qualquer caso, como pode ser visto pela comparação dos resultados das diferentes situações.

Neste mesmo caminho, um contraponto que é importante destacar são os casamentos que originalmente são classificados como Situação 4 (situação que considera casamentos como errados). Eles podem ser avaliados como verdadeiros em virtude de seu valor de similaridade final, como observamos na tabela da situação 4 que contém casamentos com os valores máximos acima da média da Situação 3. Isso acontece quando temos todos os atributos iguais e ordem de autoria e nome do pesquisador diferentes. Porém quando somados, os dados têm a capacidade de ficar a um nível acima.

Quadro 7 – Resultados dos grupos de pesquisa referente à Situação 1.

Situação 1 Pareamentos que possuem nomes iguais.					
GRUPO	**MÉDIA**	**DESVIO PADRÃO**	**VARIÂNCIA**	**VALOR MÍNIMO**	**VALOR MÁXIMO**
A	9,66	0,73	0,54	7,95	10
B	9,35	0,93	0,87	8	10
C	9,77	0,63	0,4	7,94	10

Fonte: Peil et al. (2015).

Quadro 8 – Resultados dos grupos de pesquisa referente à Situação 2.

Situação 2 Pareamentos que possuem valor >= ao nivel de trigram e ordem de coautoria correta.					
GRUPO	**MÉDIA**	**DESVIO PADRÃO**	**VARIÂNCIA**	**VALOR MÍNIMO**	**VALOR MÁXIMO**
A	7,42	0,89	0,79	5,68	9,53
B	7,14	0,91	0,83	5,82	9,2
C	7,21	0,77	0,59	5,74	9,68

Fonte: Peil et al. (2015).

Quadro 9 – Resultados dos grupos de pesquisa referente à Situação 3.

Situação 3 Pareamentos que possuem valor >= ao nível de trirgram e ordem de coautoria errado.					
GRUPO	**MÉDIA**	**DESVIO PADRÃO**	**VARIÂNCIA**	**VALOR MÍNIMO**	**VALOR MÁXIMO**
A	5,4	0,88	0,77	3,67	7,29
B	5,31	0,86	0,75	3,82	6,84
C	5,21	0,8	0,64	3,75	7,3

Fonte: Peil et al. (2015).

Quadro 10 – Resultados dos grupos de pesquisa referente à Situação 4.

Situação 4 Pareamentos que possuem valor < que o αNome e ordem de coautoria é correto.					
GRUPO	**MÉDIA**	**DESVIO PADRÃO**	**VARIÂNCIA**	**VALOR MÍNIMO**	**VALOR MÁXIMO**
A	4,35	0,52	0,27	3,9	5,63
B	4,06	0,16	0,02	4	5,57
C	4,27	0,51	0,26	3,94	5,71

Fonte: Peil et al. (2015).

Quadro 11 – Resultados dos grupos de pesquisa referente à Situação 5.

Situação 5 Pareamentos que possuem valor < αNome e ordem de coautoria errado.					
GRUPO	**MÉDIA**	**DESVIO PADRÃO**	**VARIÂNCIA**	**VALOR MÍNIMO**	**VALOR MÁXIMO**
A	2,08	0,19	0,039	1,93	3,65
B	2,07	0,18	0,03	1,95	3,57
C	2,07	0,2	0,04	1,94	3,67

Fonte: Peil et al. (2015).

4 CONSIDERAÇÕES FINAIS

Este artigo apresentou uma abordagem para alinhamento de nomes de coautores de publicações científicas, bem como sua aplicação e validação em cima de um conjunto de dados reais. Após validação do processo, os resultados são considerados satisfatórios. Os valores de similaridade de *5.2*, a *5.4* demonstram ser suficientes para representar a grande maioria dos casamentos corretos e podem apoiar o processo de de-duplicação de publicações, a identificação de redes de relacionamento científicos e outras aplicações. Devido a baixa quantidade de informações que temos com relação aos coautores nos currículos Lattes, não há como garantir exatidão nas descobertas dos alinhamentos.

A condução desta pesquisa foi motivada pela criação de um sistema Web, chamado Cientum, que apoia a tomada de decisão através da análise de desempenho de programas de pós-graduação com base nas produções de seus pesquisadores (através dos Currículos Lattes). A de-duplicação automática neste cenário evita a necessidade da identificação de duplicatas de forma manual, como é feita atualmente pelos coordenadores de pós-graduação.

Para os trabalhos futuros como ponto inicial é interessante reforçar o processo de validação, utilizando mais dados de outros grupos de pesquisa e instituições, de forma também e realizar rodízios nos conjuntos de validação. Também visamos melhorar a clusterização, estudando e testando outras técnicas que possam melhorar a abordagem proposta e reduzir, cada vez mais, a necessidade de comparações, melhorando a performance, Além do uso dos atributos de conferências e veículos de publicações, para agregar mais um nível de clusterização, e a possibilidade de desenvolvimento de novas funcionalidades no sistema Cientum.

REFERÊNCIAS

Al-Mubaid, H. and Chen, P. (2006). Biomedical term disambiguation: An application to gene-protein name disambiguation. In Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on, pages 606–612.

Álvarez, A.C.(2007). Extração de informação de artigos científicos: uma abordagem baseada em induçnao d regras de etiquetagem. PhD thesis, Universidade de São Paulo.

Brauner, D. F. (2008). Alinhamento de esquemas baseado em instâncias. PhD thesis, PUC-Rio.

Brauner, D. F., Casanova, M.A., e Milidiú, R. L. 2007. Towards gazetter integration though an instance-based thesauri mapping approach. In Advances in Geoinformatics, pages 235-245. Springer.

Chaudhuri, S., Ganjam, K., Ganti, V., and Motwani, R. (2003). Robust and efficient fuzzy match for online data cleaning. In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pages 313–324. ACM.

Cota, R. G., Gonçalves, M. A., and Laender, A. H. (2007). A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. In SBBD, pages 20–34. Citeseer.

Elmagarmid, A., Ipeirotis, P., and Verykios, V. (2007). Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on, 19(1):1–16.

Hernández, M. A. and Stolfo, S. J. (1995). The merge/purge problem for large databases. In ACM SIGMOD Record, volume 24, pages 127–138. ACM.

Kalashnikov, D. V. and Mehrotra, S. (2006). Domain-independent data cleaning via anal- ysis of entity-relationship graph. ACM Trans. Database Syst., 31(2):716–767.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707–710.

Levin, F. H. (2010). Desambiguação de autores em bibliotecas digitais utilizando redes sociais e programação genética. Master's thesis.

Mena-Chalco, J. P., Junior, C., and Marcondes, R. (2009). Scriptlattes: an open-source knowledge extraction system from the lattes platform. Journal of the Brazilian Com- puter Society, 15(4):31–39.

Peil, A. G., Brauner, D. F., e De Araújo, R. M. (2015). Um processo para apoiar a desambiguacnao de nomes em Currículos Lattes baseado em web semântica. XXIV Congresso de Iniciação Científica da UFPel

Peil, A. G., Brauner, D. F., e De Araújo, R. M. 2015. Alinhamento de nomes de Coautores em produções científicas de Currículos Lattes de pesquisadores, Monografia, Universidade Federal de Pelotas.

Rademaker, A. and Haeusler, E. H. (2013). Semantic lattes and vivo project. In Proceed- ings of VIVO 2013, St. Louis, MO.

Sun, X., Kaur, J., Possamai, L., and Menczer, F. (2011). Detecting ambiguous author names in crowdsourced scholarly data. In Privacy, Security, Risk and Trust (PAS- SAT) and 2011 IEEE Third Inernational Conference on Social Computing (Social- Com), 2011 IEEE Third International Conference on, pages 568–571. IEEE.

Torvik, V. I. and Smalheiser, N. R. (2009). Author name disambiguation in medline. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(3):11.

WC3 (2003). Rdf - resource description framework.

Weis, M. and Naumann, F. (2006). Detecting duplicates in complex xml data. In Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on, pages 109–109.