Heurísticas para desambiguação incremental de nomes de autores em referências bibliográficas

Alan Filipe Santana[1], **Marcos André Gonçalves**[1] **(Orientador)**

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais – Belo Horizonte, MG – Brasil

{alanfs, mgoncalv}@dcc.ufmg.br

1. Contextualização e Motivação

Nos últimos anos é possível observar um aumento exponencial da produção global de publicações científicas. Com o desenvolvimento da *World Wide Web* (WWW), grande parte destas publicações ficam acessíveis através da Internet. Nesse cenário, Bibliotecas Digitais (BD), como DBLP[1], MEDLINE[2] e BDBComp[3], possuem grande importância ao prover serviços que facilitam o acesso a publicações relevantes pela comunidade acadêmica, além de possibilitar pesquisas e análises relacionadas com redes de colaboração, tendências, cobertura de tópicos e impacto de publicações.

Um dos principais problemas que prejudicam a qualidade dos serviços de recuperação de informações bibliográficas fornecidos palas BDs é a ambiguidade de nomes de autores. Este problema ocorre quando dois ou mais autores compartilham um mesmo nome (nomes homônimos) ou quando um autor utiliza diferentes nomes em suas referências bibliográficas (nomes sinônimos). Os desafios ao lidar com este problema têm levado ao desenvolvimento de inúmeros métodos de desambiguação [Ferreira et al. 2012]. Dentre as estratégias possíveis, existem esforços para o desenvolvimento de identificadores únicos globais e métodos automáticos. Serviços de associação de identificadores a pesquisadores como a Open Researcher and Contributor ID (ORCID)[4] dependem da colaboração voluntária e ativa de pesquisadores e autores, o que pode ser improvável de se conseguir dentro de um intervalo de tempo curto e em âmbito global. Portanto, nos últimos anos, uma grande atenção tem sido dada ao desenvolvimento de métodos automáticos de desambiguação para serem aplicados em BDs [Ferreira et al. 2012].

Métodos automáticos de desambiguação normalmente tentam resolver o problema agrupando referências de um mesmo autor baseado em medidas de similaridades entre seus atributos ou, diretamente, associando uma citação a um determinado autor. Historicamente, métodos supervisionados têm, empiricamente, produzidos os melhores resultados [Ferreira et al. 2012]. Entretanto, ao depender de dados de treinamento, esses métodos podem não ser adequados em situações reais nas quais novos nomes ambíguos, ausentes no conjunto de treinamento, aparecem todo o tempo, além de desconsiderar as mudanças que comumente ocorrem nos perfis de publicação dos autores. Mesmos os métodos de desambiguação baseados em técnicas tradicionais de clusterização não são práticos ao considerar as características de uma BD real [Carvalho et al. 2011].

[1] http://dblp.uni-trier.de/
[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://www.lbd.dcc.ufmg.br/bdbcomp/
[4] http://orcid.org

A fim de lidar com os desafios relativos a evolução das BDs, alguns métodos utilizaram heurísticas, combinando-as com abordagens supervisionadas [Veloso et al. 2012, Ferreira et al. 2014] ou com o objetivo de se obter métodos incrementais de desambiguação [Carvalho et al. 2011, Esperidião et al. 2014]. Estes últimos visam desambiguar as referências no momento em que elas são inseridas na BD, considerando um conjunto de regras ou heurísticas para identificar a presença de um novo autor e a fragmentação de grupos de citações (ou *clusters*), portanto, são potencialmente mais eficientes. Apesar do avanço obtido com estas abordagens, alguns desafios ainda estão em aberto, especialmente os relacionados com as seguintes características: (i) presença de poucas informações nas citações, (ii) tolerância a erros, (iii) eficiência (baixa complexidade de tempo), (iv) tolerância a mudanças nos perfis de publicação dos autores e; (v) a constante inclusão de novos autores [Ferreira et al. 2012].

Essa dissertação teve como objetivo o desenvolvimento de um novo método de desambiguação de nomes autores em referências bibliográficas considerando, concomitantemente, todos os desafios encontrados em uma BD real listados acima. Particularmente, focamos no problema de *disambiguação incremental*, onde apenas as novas entradas incluídas em uma BD são desambiguadas, ao invés de aplicar ao processo de desambiguação à toda BD periodicamente, como é comumente proposto. Isso é um aspecto essencial para garantir a escalabilidade do processo, além de preservar eventuais correções manuais que porventura tenham sido efetuadas. Apesar da importância do problema, pouquíssimos trabalhos na literatura têm focado nele até bem pouco tempo atrás (apenas recentemente trabalhos têm começado a surgir, e.g., [Carvalho et al. 2011, Esperidião et al. 2014]), o que também demonstra o caráter inovador e o impacto prático de nossas soluções.

2. Principais Contribuições

As principais contribuições deste dissertação incluem:

- O desenvolvimento de um novo método de desambiguação incremental, baseado em heurísticas específicas do domínio do problema, altamente eficiente e efetivo, capaz de ser utilizado de forma supervisionada ou não.
- Análise de uma estratégia de incorporação de características baseadas em coocorrência de palavras na resolução da tarefa de desambiguação.
- Avaliação do método proposto utilizando coleções reais extraídas a partir da DBLP e BDBComp e comparação dos resultados com vários métodos supervisionados e não-supervisionados encontrados na literatura.
- Avaliação do método proposto em cenários que simulam a evolução de bibliotecas digitais durante um determinado período de tempo, utilizando coleções reais e sintéticas, e comparação dos resultados com os dois únicos métodos incrementais encontrados na literatura.

Uma parte dos trabalhos desenvolvidos foi publicada e apresentada na *ACM/IEEE Joint Conference on Digital Libraries* (JCDL) [Santana et al. 2014] (*Qualis* A2), a principal conferência mundial na área. Uma versão estendida desse trabalho foi convidada para uma edição especial dos melhores artigos da conferência e aceita para publicação, após nova revisão, no periódico *International Journal on Digital Libraries* (IJDL, *Qualis* B3) [Santana et al. 2015] . Além dessas publicações, um novo artigo abordando

a versão mais recente do método proposto foi aceita para publicação no periódico *Journal of the Association for Information Science and Technology* (JASIST, *Qualis* A1) [Santana et al. 2016], o principal periódico da área.

3. Desambiguação Incremental baseada no *Cluster* mais Similar

O método proposto na dissertação consiste basicamente em três fases: (i) seleção de *clusters* (conjuntos de citações bibliográficas) candidatos, (ii) cálculo das similaridades entre cada citação e *cluster* candidato e (iii) atualização do conjunto de treinamento. A última fase é composta pelas etapas: (a) identificação de novos autores, (b) atualização de associações duvidosas, e (c) identificação de *clusters* fragmentados. Cada fase explora heurísticas específicas do domínio com o objetivo de criar automaticamente um conjunto de treinamento utilizado para definir as associações entre referências e autores. As etapas de cada fase são detalhadas no Capítulo 3 da dissertação e resumidas a seguir.

Seleção de *clusters* candidatos: Dada uma citação de teste c_k e uma referência ambígua c_k^a, a primeira fase do método proposto consiste na seleção dos *clusters* que possuem pelo menos um nome compatível com c_k^a utilizando o algoritmo de comparação de nomes desenvolvido por [Oliveira 2005], chamado de Comparação de Fragmentos.

Cálculo das similaridades entre cada citação e *cluster* candidato: Cada termo encontrado na citação fornece uma evidência da associação entre uma referência da citação e um autor. A força desta evidência varia conforme o atributo ao qual o termo pertence e a sua capacidade discriminativa. Com base nestas observações foi escolhido utilizar uma função de similaridade baseada na soma ponderada das similaridades entre cada atributo da citação, que possui a referência ambígua, e das citações de cada *cluster* selecionado na fase anterior. Para o cálculo das similaridades entre os atributos foi utilizado uma função que pondera os termos da citação de acordo com algumas heurísticas: (i) cálculo da capacidade discriminativa do termo a partir do número de autores que o utilizam, (ii) estimativa da probabilidade condicional da referência pertencer a um dado autor dada a ocorrência do termo no conjunto de treinamento e (iii) estimativa da distribuição do termo dentro de um determinado *cluster*.

Identificação de novos autores: após o cálculo das similaridades, se o maior valor encontrado for maior ou igual a um limite γ, a citação é associada ao *cluster* correspondente, caso o contrário, um novo cluster é criado para representar o autor da referência.

Atualização de associações duvidosas: sempre que uma citação é associada a um *cluster* do conjunto de treinamento, uma métrica que estima a confiança da classificação é calculada. Se o valor dessa métrica for menor que γ, a citação é incluída em um conjunto $\mathcal{E}$ para futuras reclassificações. Caso o contrário, todas as citações presentes no conjunto $\mathcal{E}$, que compartilham pelo menos um termo com a citação classificada, são reclassificadas no intuito de corrigir possíveis erros durante o processo de desambiguação. A métrica de confiança desenvolvida utiliza as similaridades obtidas na segunda fase do algoritmo e retorna um valor entre 0 e o maior valor de similaridade encontrado.

Identificação de *clusters* fragmentados: durante a desambiguação, os grupos de citações dos autores que trabalham em diferentes linhas de pesquisa podem ser fragmentados, mesmo se forem utilizados pequenos valores para o parâmetro γ. Para lidar com esse problema, sempre que uma nova citação é incluída no conjunto de treinamento e sua

Tabela 1. Valores médios da métrica K obtidos pelos métodos supervisionados.

Método	Coleção		
	DBLP	BDBComp	KISTI
DICS	$\mathbf{0{,}919} \pm 0{,}026$	$\mathbf{0{,}956} \pm 0{,}024$	$\mathbf{0{,}952} \pm 0{,}002$
SLAND	$0{,}878 \pm 0{,}027$	$0{,}882 \pm 0{,}031$	$0{,}927 \pm 0{,}002$
Cosine	$0{,}884 \pm 0{,}028$	$0{,}746 \pm 0{,}041$	$0{,}883 \pm 0{,}003$
SVM	$0{,}777 \pm 0{,}038$	$0{,}579 \pm 0{,}042$	$0{,}797 \pm 0{,}004$
NB	$0{,}711 \pm 0{,}045$	$0{,}537 \pm 0{,}067$	$0{,}768 \pm 0{,}005$

classificação é considerada confiável, são calculadas as similaridades entre o grupo da citação e todos os outros grupos que compartilham pelo menos um termo em comum com a citação. Se essa similaridade for maior do que γ, os *clusters* são unidos. Para comparar dois *clusters* foi utilizada uma função similar a que compara uma citação com uma *cluster*, com a diferença na utilização de procedimento de normalização dos valores a partir do número de citações do menor *cluster*.

Para aumentar a praticidade e eficiência do método proposto, foram desenvolvidas heurísticas para a configuração automática do método com ou sem a presença de um conjunto de treinamento (veja as Seções 3.2 e 3.3 para mais detalhes). O algoritmo desenvolvido possui baixa complexidade de tempo, quando comparado com outras abordagens tradicionais, conforme mostrado na Seção 3.5 da dissertação.

4. Avaliação Experimental

Para avaliar o desempenho do método desenvolvido foram utilizados cinco métodos baseados em associação de autores - SAND [Ferreira et al. 2014], SLAND [Veloso et al. 2012], Cosine [Lee et al. 2005], SVM [Han et al. 2004] e NB [Han et al. 2004] -, dois métodos baseados em agrupamentos de autores - LAVSM-DBSCAN [Huang et al. 2006] e HHC [Cota et al. 2010] - e dois métodos incrementais - INDi [Carvalho et al. 2011] e MINDi [Esperidião et al. 2014]. Os experimentos foram realizados utilizando duas coleções extraídas a partir da DBLP (DBLP e KISTI), uma a partir da BDBComp e quatro coleções sintéticas. Para comparar os resultados obtidos pelos métodos de desambiguação, foram utilizadas duas métricas: a métrica K e a *pairwise* F1. A seguir são destacados os principais resultados obtidos durante a avaliação do método proposto, chamado de DICS, em diferentes cenários de aplicação.

Comparação com métodos supervisionados: Para avaliação dos métodos supervisionados foram utilizadas as coleções DBLP, BDBComp e KISTI. Os grupos ambíguos de cada coleção foram divididos em conjuntos de treino e teste, cada um com 50% do número total de citações. Esta divisão foi realizada de maneira aleatória e repetida 10 vezes. A Tabela 1 mostra os resultados obtidos em termos da métrica K. O método DICS obteve os melhores valores em todas as coleções com apenas um empate estatístico na coleção DBLP com o método Cosine em relação a métrica pF1. Em comparação com os melhores *baselines*, considerando a métrica K, o ganhos foram de 3,9%, 8,4% e 2,7% nas coleções DBLP, BDBComp e KIST, respectivamente. Em relação a métrica pF1, o maior ganho médio foi de 28% na coleção BDBComp.

Comparação com métodos não supervisionados: Para avaliação dos métodos não supervisionados foram utilizadas as mesmas divisões de teste usadas nos experimentos com os métodos supervisionados. A Tabela 2 mostra os resultados obtidos em termos

Tabela 2. Valores médios da métrica K dos métodos não supervisionados.

Método	Coleção		
	DBLP	BDBComp	KISTI
DICS	$\mathbf{0{,}791} \pm 0{,}059$	$\mathbf{0{,}944} \pm 0{,}025$	$\mathbf{0{,}942} \pm 0{,}003$
SAND	$0{,}674 \pm 0{,}091$	$\mathbf{0{,}942} \pm 0{,}022$	$0{,}892 \pm 0{,}003$
HHC	$0{,}692 \pm 0{,}084$	$\mathbf{0{,}937} \pm 0{,}021$	$0{,}862 \pm 0{,}003$
LASVM-DBSCAN	$0{,}479 \pm 0{,}097$	$\mathbf{0{,}883} \pm 0{,}042$	$0{,}858 \pm 0{,}004$

da métrica K. O método DICS obteve os melhores resultados médios nas coleções DBLP e KISTI. Na coleção BDBComp houve empate estatístico entre DICS e os métodos SAND e HHC. Os ganhos obtidos em relação ao melhor *baseline* foram de 14% e 5,5% nas coleções DBLP e KISTI respectivamente. Comparando com os resultados obtidos com a utilização do conjunto de treinamento, a queda de desempenho foi de 13,9% na DBLP e menos de 2% na coleções BDBComp e KISTI.

Comparação com métodos incrementais: Para a avaliação dos métodos incrementais, foram utilizadas as coleções BDBComp e KISTI ordenadas em ordem cronológica e quatro coleções sintéticas que simulam cenários onde novos autores são inseridos no repositório da BD e onde os autores alteram seus perfis de publicação. Em relação à métrica K, DICS obteve os maiores valores com ganhos de até 8,7% e 50% comparando com INDi nas coleções BDBComp e KISTI, respectivamente. Analisando os resultados em termos de coesão (PMA) e pureza (PMC), DICS foi por volta de 7,3% menos efetivo em relação à métrica PMC quando comparado com o método INDi, mas cerca de 26% e 143% mais efetivo considerando a métrica PMA nas coleções BDBComp e KISTI, respectivamente. Nas coleções sintéticas DICS, superou todos os baselines em termos de coesão e equilíbrio entre fragmentação e pureza. Nos cenários simulando a introdução de novos autores, o ganho, comparando com o melhor *baseline*, alcançou 4,5% em relação à métrica K. Nos cenários simulando mudanças nos perfil de publicação dos autores, o método proposto obteve ganhos de até 19,8% comparado com o método MINDi.

Além destas avaliações, foram realizadas diversos outros experimentos que avaliaram: (i) a utilização da estratégia de incorporação de características baseadas na coocorrência de palavras proposto por [Figueiredo et al. 2011], (ii) o impacto de cada componente do algoritmo, (iii) a sensibilidade do método DICS aos valores dos seus parâmetros e (iv) a análise do tempo de execução.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi proposto um novo método incremental de desambiguação baseado em heurísticas capaz de criar e atualizar automaticamente um conjunto de treinamento utilizado para determinar os autores de cada citação. Foram propostos procedimentos para realizar a configuração automática dos parâmetros com e sem a presença de dados de treinamento. Foi realizada uma extensa avaliação experimental, utilizando coleções reais e sintéticas, a fim de avaliar o desempenho do método quando este é aplicado de forma supervisionada, não supervisionada e incremental. Em todos os cenários, o método proposto superou todos os *baselines*, com ganhos em quase todos os grupos ambíguos. Também foram apresentadas: a avaliação de uma estratégia de incorporação de características baseadas em coocorrência de palavras, uma análise das capacidades do método, uma análise de sensibilidade aos valores dos parâmetros e uma análise da complexidade do algoritmo.

No futuro, pretende-se: (i) avaliar estratégias para alteração automática dos valores dos parâmetros à medida que o método atualiza o conjunto de treinamento; (ii) avaliar a inclusão de novos atributos, quando disponíveis, como ano da publicação e endereço dos autores; e (iii) explorar estratégias de *relevance feedback* para permitir melhorias no modelo de desambiguação a partir da interação com um administrador da BD.

6. Agradecimentos

Esse trabalho foi parcialmente financiado pelo projeto INWeb (MCT/CNPq 573871/2008- 6) e por financiamentos individuais de CNPq, CAPES e FAPEMIG.

Referências

Carvalho, A. P., Ferreira, A. A., Laender, A. H. F., and Gonçalves, M. A. (2011). Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries. *J. of Inf. and Data Manage.*, 2(3):289–304.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., and Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *J. of the American Soc. for Info. Sci. and Tech.*, 61(9):1853–1870.

Esperidião, L. V. B., Ferreira, A. A., Laender, A. H. F., Gomes, D. M., Tavares, A. I., and Assis, G. T. (2014). Reducing Fragmentation in Incremental Author Name Disambiguation. *JIDM*, 5(3).

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, 41(2):15–26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., and Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *JASIST*, 65(6):1257–1278.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., and Meira Jr., W. (2011). Word Co-occurrence Features for Text Classification. *Inf. Syst.*, 36(5):843–858.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two Supervised Learning Approaches for Name Disambiguation in Author Citations. In *Proc. of the 4th ACM/IEEE-CS JCDL*, JCDL '04, pages 296–305, New York, NY, USA. ACM.

Huang, J., Ertekin, S., and Giles, C. L. (2006). Efficient Name Disambiguation for Large-scale Databases. In *Proc. of the 10th European Conf. on PKDD*, pages 536–544, Berlin, Heidelberg. Springer-Verlag.

Lee, D., On, B.-W., Kang, J., and Park, S. (2005). Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries. In *Proc. of the 2nd int. workshop on Info. Quality in Info. Syst.*, pages 69–76.

Oliveira, J. W. A. (2005). Uma estratégia para remoção de ambiguidades na identificação de autoria de objetos bibliográficos. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Santana, A. F., Goncalves, M. A., Laender, A. H. F., and Ferreira, A. A. (2014). Combining domain-specific heuristics for author name disambiguation. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, pages 173–182.

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., and Ferreira, A. A. (2015). On the combination of domain-specific heuristics for author name disambiguation: the nearest cluster method. *International Journal on Digital Libraries*, 16(3):229–246.

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., and Ferreira, A. A. (2016). Incremental author name disambiguation by exploiting domain-specific heuristics. *JASIST.*, (Aceito para publicação).

Veloso, A., Ferreira, A. A., Gonçalves, M. A., Laender, A. H. F., and Meira, Jr., W. (2012). Cost-effective On-demand Associative Author Name Disambiguation. *Inf. Proc. & Manage.*, 48(4):680–697.