UDRB: Uma Nova Heurística Eficaz para Deduplicação de Referências Bibliográficas

Sérgio Canuto[1], Guilherme Dal Bianco[2], Marcos André Gonçalves[1], Jussara Almeida[1], Thierson Couto[3]

[1] Universidade Federal de Minas Gerais, Brasil
{sergiodaniel, mgoncalv, jussara}@dcc.ufmg.br
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
gbianco@inf.ufrgs.br
[3] Universidade Federal de Goiás, Goiânia, GO, Brasil
thierson@inf.ufg.br

Abstract. Publicações científicas normalmente contêm referências bibliográficas a trabalhos anteriores. Tais referências são usadas como fonte de informação para bibliotecas digitais, contribuindo com recursos de busca, navegação e estimativa de qualidade das obras. Neste contexto, frequentemente ocorre um problema que consiste em identificar se duas referências representam uma mesma publicação, conhecido como deduplicação de referências bibliográficas (DRB). Soluções para DRB podem ser divididas em supervisionadas (dependem de um conjunto de treinamento) e não supervisionadas (baseados em heurísticas). Com objetivo de evitar o acentuado custo manual de criação de um conjunto de treinamento, propomos neste trabalho uma heurística não supervisionada para DRB, denominada UDRB. Os experimentos em bases reais mostraram que a heurística proposta alcançou ganhos de mais de 7% em relação ao método não supervisionado estado-da-arte, e eficácia similar as de métodos supervisionados na maioria dos casos, sem a necessidade da dispendiosa tarefa de rotulação manual.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

Keywords: Deduplicação de Referências, Desambiguação

1. INTRODUÇÃO

O surgimento da Web tem permitido a coleta de uma grande quantidade de informações bibliográficas sobre artigos científicos. Entre as várias informações disponíveis, destacam-se as referências bibliográficas que aparecem geralmente nas últimas seções dos artigos. Quando consideradas coletivamente, as referências bibliográficas permitem mensurar a importância de uma determinada obra, por exemplo, por meio da contagem do número de trabalhos de uma coleção que possuem referência bibliográfica a tal obra. Alguns índices de reputação de obras científicas baseiam-se nesta contagem, como é o caso do *fator de impacto*, muito utilizado para medir a importância de revistas indexadas no meio científico[1].

As métricas usadas para avaliar o impacto de uma publicação são de grande importância prática, contudo, existe um problema que antecede o seu cálculo, o qual consiste em identificar quais referências bibliográficas distintas se referem a uma mesma obra científica. No contexto de bibliotecas digitais, esse problema é denominado *deduplicação de referências bibliográficas* - DRB.

É comum encontrarmos em uma biblioteca digital centenas de milhares de referências bibliográficas e, nesse contexto, uma solução manual não escala. É necessário o desenvolvimento de soluções automáticas que possam ser executadas por computadores. A dificuldade de se conseguir soluções automáticas para DRB se deve às seguintes razões: (i) referências a uma mesma obra podem estar escritas de formas muito diferentes entre si, normalmente por questões de estilo. Por exemplo, abreviações de iniciais nos nomes de autores e a presença (ou ausência) de numeração de páginas são fatores que podem diferenciar referências que, na verdade, correspondem à mesma obra; (ii) referências a publicações distintas mas que possuem o texto muito parecido. Um exemplo típico ocorre quando um autor publica uma versão preliminar de seu trabalho em uma conferência e posteriormente publica uma versão estendida do mesmo trabalho em um periódico.

[1] http://thomsonreuters.com/journal-citation-reports

Para lidar com esses problemas, dois grupos de métodos para DRB foram propostos: *supervisionados* e *não-supervisionados*. Os métodos DRB supervisionados [Bilenko and Mooney 2003; Borges et al. 2012] dependem de um conjunto de treinamento informativo para a identificação dos padrões presentes na base de dados. Se o conjunto de treinamento não for suficientemente informativo, o método provavelmente não atingirá a eficácia esperada. Já os métodos não-supervisionados [Borges et al. 2011; McCallum et al. 2000; Lawrence et al. 1999] utilizam heurísticas para identificar os padrões de duplicatas, como por exemplo, medidas de similaridade entre os campos autores, título e ano de duas referências [Borges et al. 2011]. A principal vantagem das técnicas não supervisionadas é a capacidade de atingir uma eficácia muitas vezes comparável a de métodos supervisionados sem a necessidade da dispendiosa tarefa de criação do conjunto de treinamento.

Neste trabalho propomos um método não supervisionado para DRB: o UDRB. Tal abordagem adiciona um conjunto de melhorias na heurística não-supervisionada proposta em [Borges et al. 2011], considerada estado-da-arte, aprimorando assim sua eficácia. Nossa experimentação mostrou que o UDRB atinge uma eficácia equivalente a dos melhores métodos supervisionados recentemente descritos em [Borges et al. 2012], sem a necessidade da criação de uma amostra manualmente rotulada. Mostramos também que a estratégia UDRB sempre é competitiva, mesmo quando as estratégias supervisionadas são aprimoradas com um novo conjunto de atributos aqui propostos. Portanto, a segunda contribuição do trabalho é a melhoria nos métodos supervisionados de [Borges et al. 2012] através da extensão do conjunto de seus atributos com diversas medidas de similaridade, como Levenshtein, Jaro-Winkler, Jaccard, *softTFIDF* [Cohen et al. 2003], bem como medidas utilizadas em trabalhos de deduplicação anteriores [McCallum et al. 2000; Lawrence et al. 1999]. Com essas medidas, é possível capturar evidências de similaridade que consideram abreviações, erros tipográficos, inversões de termos, e especificidades de cada campo das referências.

A última contribuição deste trabalho consiste na análise do impacto na eficácia da deduplicação de quatro fatores normalmente considerados em métodos de deduplicação: a segmentação das referências em campos [Isaac Councill 2008], a blocagem [McCallum et al. 2000] (que reduz a quantidade de comparações entre pares), a identificação de referências associadas à mesma publicação, e a geração de uma solução consistente (i.e., que lida com o problema da transitividade [Huang et al. 2006], que surge se o método classifica tanto a e b quanto b e c como correferentes, mas a e c como não correferentes).

Este artigo é organizado da seguinte forma. A Seção 2 apresenta trabalhos relacionados. A Seção 3.1 apresenta o método não supervisionado para DRB do estado-da-arte seguido pela nossa proposta. O conjunto de atributos propostos para deduplicação supervisionada é apresentado na Seção 4. Experimentos e resultados são discutidos na Seção 5, enquanto a Seção 6 conclui o artigo.

2. TRABALHOS RELACIONADOS

Até pouco tempo, as soluções para a DRB utilizavam diretamente o texto das referências no processo de deduplicação. Entretanto, esta abordagem leva a resultados que não são satisfatórios devido à grande quantidade de ruído existente nos textos. Recentemente, técnicas de extração de informação [Isaac Councill 2008] têm sido empregadas para separar o texto de cada referência bibliográfica em seus campos constituintes, tais como: autores, título, veículo de publicação, data, etc. Desse modo, as soluções para DRB podem trabalhar com campos correspondentes de duas referências bibliográficas, obtendo mais precisão no processo de deduplicação [McCallum et al. 2000].

As técnicas supervisionadas (que utilizam um conjunto de dados de treinamento) apresentam bons resultados na deduplicação de referências. Em tal abordagem, uma função de deduplicação é gerada através de estratégias como SVM treinado com atributos baseados na co-ocorrência de termos [Bilenko and Mooney 2003], bem como *Naive Bayes*, SVM, e Árvores de Decisão (AD) usando similaridades entre campos das referências como atributos [Borges et al. 2012]. Neste trabalho, aprimoramos os resultados recentes dos métodos supervisionados descritos em [Borges et al. 2012]. Para tal, empregamos abordagens como SVM, AD e Florestas Randômicas (FR), e estendemos o conjunto

de atributos explorado para incluir também métricas relacionadas com co-ocorrência de termos e similaridades entre campos das referências.

As soluções não supervisionadas [Borges et al. 2011; McCallum et al. 2000; Lawrence et al. 1999] identificam se referências estão associadas à mesma publicação através de heurísticas. A melhor heurística proposta em [Lawrence et al. 1999] consiste em quantificar termos em comum entre referências, ignorando a estrutura de campos da referência. Para utilizar essa estrutura de campos, [McCallum et al. 2000] combina linearmente as similaridades entre os campos das referências. Recentemente, o trabalho [Borges et al. 2011] propõe um conjunto de regras para combinar similaridades específicas correspondentes a cada campo das referências, apresentando resultados bastante positivos. Entretanto, tal heurística considera apenas os campos ano, autores e título, e falha ao deduplicar os casos em que algum dos campos da referência não está presente. Para superar esse problema, o método não supervisionado aqui proposto complementa o método em [Borges et al. 2011] utilizando, em adição aos campos, o texto completo das referências bibliográficas. O método proposto melhora sensivelmente a deduplicação quando os campos não são corretamente identificados ou estão ausentes.

3. HEURÍSTICAS PARA DEDUPLICAÇÃO NÃO-SUPERVISIONADA

Nesta seção são apresentados os métodos *MetadataMatch* e UDRB que correspondem, respectivamente, ao estado-da-arte para deduplicação de referências bibliográficas de modo não supervisionado, e à heurística não supervisisionada proposta.

3.1 Baseline: Deduplicação MetadataMatch

MetadataMatch propõe a utilização de um conjunto de regras, definidas de acordo com o conteúdo de cada campo, para identificar correferências de forma automática. Tal heurística considera que um pré-processamento foi otimamente realizado para segmentar em campos. Mais especificamente, o Algoritmo 1 detalha a heurística utilizando três medidas para verificar se duas referências a e b descrevem a mesma publicação. A primeira medida consiste na diferença entre o ano de publicação de tais referências, denotada por $|Year(a) - Year(b)|$. Tal diferença deve ser menor que um limiar t_Y para que a e b sejam consideradas correferentes (Linha 2). A segunda medida, denotada pela função *NameMatch*, mensura a similaridade entre os autores de a e b (Linha 3). Essa função compara as iniciais de todos autores de a com as iniciais de todos autores de b, considerando que tais iniciais podem aparecer invertidas. Por fim, o título das referências é comparado usando a medida de similaridade Levenshtein normalizada [Borges et al. 2011]. Caso essas três medidas respeitem os seus respectivos limiares t_Y, t_N e t_L, as referências a e b são consideradas correferentes (Linha 5).

Algorithm 1 Heurística MetadataMatch proposta por [Borges et al. 2011].

```
1: function MetadataMatch(a, b, t_Y, t_N, t_L)
2:     if |Year(a) − Year(b)| ≤ t_Y then
3:         if NameMatch(Authors(a), Authors(b)) ≥ t_N then
4:             if Levenshtein(Title(a), Title(b)) ≥ t_L then
5:                 return 1
6:     return 0
```

3.2 Nova Heurística: Deduplicação UDRB

A heurística de deduplicação UDRB, proposta neste trabalho, visa estender o método *MetadataMatch* ao considerar também todo o texto das referências, isso é, o texto original não segmentado. Como o pré-processamento para a extração automática das citações pode falhar, ou os campos podem nem mesmo existir no texto da referência, a heurística *MetadataMatch* pode atingir uma qualidade aquém da desejada. Neste contexto, a heurística UDRB utiliza o texto completo das referências, e considera os casos em que não é possível extrair os campos de uma referência com precisão.

O Algoritmo 2 difere do *MetadataMatch* por incluir os conectivos lógicos *OR* para que as medidas de similaridade sobre os campos ano, autores e título sejam desconsideradas caso não tenham sido extraídos. Além disso, foi incluída a medida *Jaccard* [Cohen et al. 2003] para comparar todo o texto de duas referências (Linha 5). Logo, todas as palavras presentes no texto original da referência a podem

ser comparadas as da referência b sem considerar a divisão em campos. Para efeito de comparação, foi criado um *baseline* denominado Jaccard-DRB que corresponte ao uso da Linha 5 somente.

Algorithm 2 UDRB: nova heurística de deduplicação.

```
1: function UDRB(a, b, t_Y, t_N, t_L, t_F)
2:     if |Year(a) − Year(b)| ≤ t_Y  OR  Year(a) = ∅ OR  Year(b) = ∅  then
3:         if Namematch(Authors(a), Authors(b)) ≥ t_N  OR  Authors(a) = ∅ OR  Authors(b) = ∅ then
4:             if Levenshtein(Title(a), Title(b)) ≥ t_L  OR  Title(a) = ∅ OR  Title(b) = ∅ then
5:                 if Jaccard(FullText(a), FullText(b)) ≥ t_F then
6:                     return 1
7:     return 0
```

4. ATRIBUTOS PARA DEDUPLICAÇÃO SUPERVISIONADA

Nesta seção é proposto um conjunto de atributos específicos para deduplicação de referências. Tais atributos podem ser usados com estratégias supervisionadas conhecidas de aprendizagem de máquina, como SVM, Árvores de Decisão (AD) e Florestas Randômicas (FR). Recentemente, as estratégias SVM e AD foram usadas em [Borges et al. 2012]. Entretanto, tais métodos de aprendizagem foram treinados com um conjunto reduzido de 8 atributos, destacados em negrito na Tabela I. Tais atributos consideram a quantidade de autores de cada referência, a diferença entre essas quantidades, e as medidas usadas pela heurística *MetadataMatch* (em conjunto com suas versões normalizadas).

Neste trabalho, estendemos essas medidas com um conjunto mais robusto de atributos capazes de extrair informações relevantes para estratégias supervisionadas de DRB. Vale ressaltar que não é do nosso conhecimento que o conjunto completo de atributos proposto neste trabalho tenha sido utilizado em algum trabalho anterior. Mais especificamente, a Tabela I ilustra o conjunto de atributos que propomos para complementar o conjunto usado em [Borges et al. 2012]. A medida softTFIDF [Cohen et al. 2003] foi usada na maioria dos campos, visto que foi criada para lidar com o peso TFIDF das palavras considerando erros tipográficos e abreviações. Para considerar a ordem das palavras no texto das referências, foi utilizada a estratégia *Jaccard2gram*, como sugerido em [Lawrence et al. 1999]. Essa medida fornece um índice proporcional à quantidade de pares de palavras consecutivas que aparecem em ambas referências. As medidas *Jaro-Winkler* [Cohen et al. 2003] e *SameType* foram usadas especificamente para lidar, respectivamente, com os campos título e veículo de divulgação. *Jaro-Winkler* lida com o fato de que um título pode aparecer incompleto, e *SameType* retorna 1 caso ambas referências são pertencentes ao mesmo tipo de veículo (isto é, ambas foram publicadas em um periódico, por exemplo) e 0 caso contrário.

Campo	Medida
Ano	**|Year(a)-Year(b)|**
Páginas	Levenshtein(Pages(a), Pages(b))
Instituição	softTFIDF(Instituition(a), Instituition(b))
Localização	softTFIDF(Location(a), Location(b))
Veículo	SameType(Venue(a), Venue(b)), softTFIDF(Venue(a), Venue(b))
Autores	**NameMatch(Authors(a),Authors(b)),Norm_NameMatch(Authors(a),Authors(b)),Num(Authors (a)), Num(Authors(b)), |Num(Authors(a))-Num(Authors(b))|**,softTFIDF(Authors(a),Authors(b))
Título	**Levenshtein(Title(a), Title(b)), Norm_Levenshtein(Title(a), Title(b))**, Jaccard(Title(a), Title(b)), softTFIDF(Title(a), Title(b)), Jaccard2gram(Title(a), Title(b)), Jaro-Winkler(Title(a), Title(b))
Texto todo	Jaccard(FullText(a),FullText(b)),Jacard2gram(FullText(a),FullText(b)),SoftTFIDF(FullText(a),FullText(b))

Tabela I. Conjunto de medidas de similaridade para DRB. As medidas em negrito foram usadas em [Borges et al. 2012].

5. RESULTADOS EXPERIMENTAIS

Nesta seção, avaliamos as abordagens não supervisionadas UDRB e *MetadataMatch*, bem como os métodos supervisionados SVM, AD e FR executados com os atributos de [Borges et al. 2012], denominados respectivamente, SVM-Borges, AD-Borges e FR-Borges, ou executados com o conjunto de atributos propostos, denominados SVM-DRB, AD-DRB e FR-DRB. A eficácia dos resultados foi mensurada pela medida F1 [Borges et al. 2011]. Todos os experimentos foram feitos usando o procedimento de validação cruzada em 4 partições (*folds*), e portanto, apresentamos o F1 médio destas 4 execuções. A significância estatística foi confirmada por meio de um teste-t pareado, com 95% de confiança. Foram utilizadas duas bases de dados reais: a coleção CiteSeer, com 1563 citações a 906 publicações [Lawrence et al. 1999], e a coleção Cora com 2191 citações a 305 publicações.

5.1 Parametrização

Foi adotada a implementação LIBLINEAR do classificador SVM. O parâmetro de regularização C foi escolhido entre onze valores, de 2^{-5} até 2^{15} através da estratégia de validação cruzada em 4 partições. As árvores dos métodos FR e AD foram geradas sem poda. Como sugerido na literatura para o método FR, utilizamos a quantidade de árvores $T > 100$ e a quantidade de atributos considerados a cada divisão $M = log_2 F$, em que F é a quantidade total de atributos.

Quanto a heurística *MetadataMatch*, foram testados os valores 0, 1 e 2 para o parâmetro t_Y, e valores de 0,4 a 0,7, variando de 0,05 em 0,05, para os parâmetros t_L e t_N. Para a nova heurística UDRB (e para Jaccard-DRB), foram testados também valores de 0,2 a 0,6 para o parâmetro adicional t_F. Foi feito um projeto fatorial completo com quatro replicações, considerando todas as possíveis combinações de valores considerados para os parâmetros. A Tabela II apresenta os parâmetros que levaram aos melhores resultados no treino.

Abordagem	Parâmetros
MetadataMatch	Para Cora, t_L=0,6, t_Y=1 e t_N=0,5. Para CiteSeer, t_L=0,55, t_Y=0 e t_N=0,5
UDRB	Para Cora, t_L=0,65, t_Y=1, t_N=0,5 e t_F=0,35. Para CiteSeer, t_L=0,5, t_Y=1, t_N=0,5 e t_F=0,35
Jaccard-DRB	Para as coleções CiteSeer e Cora, t_F=0,5
SVM-DRB e SVM-Borges	Para as coleções CiteSeer e Cora $C = 2$, Kernel linear
FR-DRB e FR-Borges	Para as coleções CiteSeer e Cora, $T = 500$, $M = log_2 F$
AD-DRB e AD-Borges	Para as coleções CiteSeer e Cora, árvore gerada sem poda

Tabela II. Parâmetros adotados nos experimentos.

5.2 Eficácia das Abordagens de Deduplicação

Nos experimentos apresentados na Tabela III foi realizado um pré-processamento para corrigir casos em que o campo correspondente ao ano da publicação não existe nos metadados. Para *MetadataMatch**, esse pré-processamento foi feito como descrito em [Borges et al. 2011]. Para lidar com as demais abordagens da Tabela III, propomos extrair o ano através da busca por um número entre 1950 e 2013 em todo o texto da referência. Com essa pequena modificação, *MetadataMatch* superou em 4% sua versão *MetadataMatch** com pré-processamento original.

O método UDRB supera MetadataMatch com significância estatística, com ganhos de 7% e 9% nas coleções Cora e CiteSeer, respectivamente. Tais melhorias advêm da exploração do texto completo das referências e da consideração dos casos em que campos das referências não foram devidamente segmentados. Em geral, UDRB também mostrou uma eficácia comparável (e em alguns casos superior) à dos métodos supervisionados, que exigem a dispendiosa tarefa de rotulação de pares de referências. Especificamente, UDRB apresentou ganhos de 3% a 6% sobre métodos supervisionados SVM-Borges, AD-Borges e FR-Borges na coleção Cora e empatou com os mesmos na coleção Citeseer. UDRB também empatou com SVM-DRB, AD-DRB e FR-DRB na coleção Citeseer, mas na coleção Cora os métodos AD-DRB e FR-DRB apresentaram ganhos de cerca de 3% sobre UDRB. Tais ganhos são pequenos, tendo em vista o custo adicional de rotulamento.

Abordagens da Literatura	Cora	CiteSeer	Abordagens Propostas	Cora	CiteSeer
MetadataMatch*	$83,3 \pm 2,0 \downarrow$	$84,2 \pm 2,0 \downarrow$	Jaccard-DRB	$88,2 \pm 4,0 \downarrow$	$85,9 \pm 2,6 \downarrow$
MetadataMatch	$86,9 \pm 0,9 \downarrow$	$85,0 \pm 3,1 \downarrow$	UDRB	$94,9 \pm 1,1$	$91,2 \pm 3,3$
SVM-Borges	$89,2 \pm 4,1 \downarrow$	$91,5 \pm 5,9 \updownarrow$	SVM-DRB	$94,7 \pm 2,2 \updownarrow$	$92,5 \pm 7,1 \updownarrow$
AD-Borges	$91,7 \pm 2,6 \downarrow$	$88,6 \pm 10,3 \updownarrow$	AD-DRB	$97,0 \pm 1,2 \uparrow$	$94,0 \pm 4,2 \updownarrow$
FR-Borges	$90,4 \pm 4,5 \downarrow$	$92,0 \pm 5,0 \updownarrow$	FR-DRB	$97,7 \pm 1,0 \uparrow$	$95,0 \pm 5,6 \updownarrow$

Tabela III. Eficácia média, em F1, das diferentes abordagens para deduplicação (e intervalos com 95% de confiança). $\downarrow$, $\uparrow$, $\updownarrow$ indicam, respectivamente, perdas, ganhos e empates com UDRB. *Executada com pré-processamento da literatura.

Como esperado, houve ganhos significativos dos métodos SVM-DRB, AD-DRB e FR-DRB sobre suas respectivas versões da literatura SVM-Borges, AD-Borges e FR-Borges na coleção Cora, pois os métodos da literatura não utilizam diversos dos atributos descritos na Tabela I. Na coleção CiteSeer, os ganhos não foram significativos devido à grande variação dos resultados, que reflete nos largos intervalos de confiança (cerca de três vezes maiores que os da Cora). Tal variação deve-se ao fato que a coleção não foi dividida aleatoriamente, como a Cora. Ao invés disso, as referências foram particionadas por área, o que causa diferença na distribuição dos dados.

5.3 Análise dos Fatores que Influenciam a Eficácia da Deduplicação

Com o objetivo de avaliar a importância e a interação entre quatro fatores que influenciam a eficácia da deduplicação, executamos um projeto fatorial $2^k r$ [Jain 1991] para analisar o impacto de cada um deles na variação dos dados. Cada fator pode assumir um de dois níveis, que correspondem às possíveis variações consideradas para tal fator, conforme especificado na Tabela IV.

Fator	Níveis do fator	
Segmentação	**Seg. Parscit**	Experimentos com a segmentação automática ParsCit [Isaac Councill 2008].
	Seg. Original	Experimentos usando a segmentação original da coleção.
Blocagem	**Com Bloc.**	Experimentos usando o método de blocagem *canopy* [McCallum et al. 2000].
	Sem Bloc.	Experimentos sem usar estratégia de blocagem.
Deduplicação	**MetadataMatch**	Experimentos usando *MetadataMatch* para deduplicação.
	UDRB	Experimentos usando a heurística UDRB proposta.
Transitividade	**Com Trans.**	Experimentos resolvem o problema de transitividade [Huang et al. 2006].
	Sem Trans.	Experimentos sem lidar com o problema de transitividade

Tabela IV. Variação dos fatores que impactam na eficácia da deduplicação.

De acordo com o projeto fatorial desenvolvido, cujos resultados são mostrados na Tabela V, os fatores que mais impactam a eficácia da deduplicação são a estratégia de deduplicação adotada, responsável por 60% da variação nos dados, e o método de segmentação adotado, responsável por 8% da variação. Em geral, a abordagem UDRB se saiu muito melhor que o método *MetadataMatch* para deduplicação, o que explica o papel central da estratégia usada na eficácia dos resultados. O fator segmentação também se mostrou importante, pois a utilização da segmentação automática ParsCit introduz erros que se mostraram, em geral, prejudiciais aos métodos. Além destes dois fatores primários, a interação entre a deduplicação e os métodos de blocagem usados também se mostrou importante, explicando 10% da variação nos dados. Isto se deve ao fato de que a eficácia do método *MetadataMatch* sempre é aprimorada com a utilização da estratégia de blocagem, pois a blocagem é baseada na comparação de todo texto das referências com a medida de similaridade Jaccard. Logo, a inclusão do blocagem no MetadataMatch de certa forma inclui também essa medida de similaridade no método. Os demais fatores e interações entre fatores, quando significativos, foram responsáveis por variações desprezíveis nos resultados.

		Seg. ParsCit		Seg. Original	
		Com bloc.	Sem bloc.	Com bloc.	Sem bloc.
UDRB	Com trans.	$91,7 \pm 1,1$	$92,3 \pm 1,1$	$94,9 \pm 1,1$	$94,4 \pm 1,3$
	Sem trans.	$92,0 \pm 1,0$	$91,8 \pm 1,0$	$93,8 \pm 1,2$	$93,4 \pm 1,0$
MetadataMatch	Com trans.	$88,6 \pm 3,2$	$85,5 \pm 1,5$	$90,3 \pm 4,9$	$86,9 \pm 0,9$
	Sem trans.	$90,5 \pm 1,8$	$85,6 \pm 1,2$	$91,5 \pm 3,0$	$86,3 \pm 0,9$

Tabela V. Eficácia dos fatores que impactam na deduplicação da coleção Cora (e intervalos com 95% de confiança).

6. CONCLUSÃO E TRABALHOS FUTUROS

Neste artigo propomos um novo método não-supervisionado para deduplicação de referências, que se mostrou significativamente superior à heurística de deduplicação do estado-da-arte e com eficácia equivalente aos melhores métodos supervisionados considerados. A avaliação da performance dos métodos usando diferentes estratégias de segmentação é um trabalho atualmente em andamento.

REFERÊNCIAS

Bilenko, M. and Mooney, R. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, 2003.

Borges, E. N., Becker, K., Heuser, C. A., and Galante, R. A classification-based approach for bibliographic metadata deduplication. *IADIS*, 2012.

Borges, E. N., de Carvalho, M., Galante, R., Gonçalves, M. A., and Laender, A. An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Inf. Process. Manage.*, 2011.

Cohen, W. W., Ravikumar, P. D., and Fienberg, S. E. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, 2003.

Huang, J., Ertekin, S., and Giles, C. L. Efficient name disambiguation for large-scale databases. In *PKDD*, 2006.

Isaac Councill, C. Lee Giles, K. Parscit: An open-source crf reference string parsing package. In *LREC*, 2008.

Jain, R. K. *The Art of Computer Systems Performance Analysis*. Wiley, 1991.

Lawrence, S., Giles, C. L., and Bollacker, K. D. Autonomous citation matching. In *AGENTS*, 1999.

McCallum, A., Nigam, K., and Ungar, L. H. Efficient clustering of high-dimensional data sets with application to reference matching. In *KDD*, 2000.