

Um método para identificação de um conjunto representativo de citações bibliográficas para remoção de ambiguidade de nomes de autores de artigos científicos

Herculano Gripp Neto

Departamento de Computação - Universidade Federal de Ouro Preto

22 de janeiro de 2013

Sumário

- 1 Introdução
 - Bibliotecas Digitais
 - Ambiguidade de Nomes
 - Justificativa
 - Objetivos
- 2 Fundamentação Teórica
 - Métricas de similaridade entre cadeias de caracteres
 - Métricas de Avaliação
 - Baseline - INDi
- 3 Método Proposto
 - Escolha de Registro Representativo
 - Remoção de Registro Representativo
- 4 Avaliação Experimental
- 5 Conclusão

- O problema de ambiguidade de nomes é observado em diversos contextos.
- Afeta principalmente sistemas computacionais.
- Será abordado o problema de ambiguidade de nomes de autores em registros de artigos científicos.
- Um artigo científico em uma biblioteca digital é representado pelos seus meta-dados.

Bibliotecas Digitais

- São sistemas de informação complexos, que são projetados para um público específico, possuem um conjunto grande de objetos digitais e seus meta-dados, várias estruturas organizacionais e fornecem diversos serviços para manter e acessar esses objetos digitais [Gonçalves et al., 2004].
- São fontes massivas de informação para diversos segmentos.
- Exemplos: DBLP, BDBComp(UFMG), NUPILL(UFSC), Brasileira (USP), MEDLINE.

Ambiguidade de Nomes

- Segundo Lee et al. [2005] podemos dividi-lo em dois sub problemas: split citation e mixed citation.
- Split Citation (SC)
 - Variação na representação do nome do autor.
 - Publicações de um mesmo autor podem estar divididas.
- Mixed Citation (MC)
 - Diferentes autores compartilham o mesmo nome ou a mesma variação de nome.
 - Publicações aparecem como se pertencessem a um mesmo autor.

Justificativa

- Afeta busca por artigos de um determinado autor.
- Afeta a análise de padrões de colaboração em redes sociais.
- Afeta análise de qualidade de impacto das publicações.

Objetivos

- Objetivo Geral

- Manter uma coleção de registros de artigos científicos livre de ambiguidade usando na desambiguação menor número de registros.

- Objetivos Específicos

- Fazer uma revisão bibliográfica sobre métodos de remoção de ambiguidade.
- Propor um método que identifique os registros mais representativos de um grupo de autores.
- Avaliar os resultados do método proposto aplicado a um método da literatura.

Métricas de similaridade entre cadeias de caracteres

- Permitem avaliar quantitativamente a semelhança entre duas cadeias de caracteres.
- Retornam valores entre 0 e 1.
- Distância de Levenshtein [1966].

$$lev(S, T) = 1 - \frac{ed(S, T)}{\min(S, T)}$$

- Comparação por Fragmentos [French et al., 2000].
 - Avalia um a um cada fragmento de duas cadeias.
 - Utiliza o algoritmo de Levenshtein para esta avaliação.

Métricas de Avaliação

- A métrica K [Lapidot, 2002], determina o equilíbrio entre duas métricas específicas de agrupamento: o PMC e o PMA.
- Métrica PMC - Pureza Média do Grupo

$$PMC = \frac{1}{N} \sum_{i=1}^q \sum_{j=1}^R \frac{n_{ij}^2}{n_i}$$

- Métrica PMA - Pureza Média do Autor

$$PMA = \frac{1}{N} \sum_{j=1}^R \sum_{i=1}^q \frac{n_{ij}^2}{n_j}$$

- Métrica K

$$k = \sqrt{PMC * PMA}$$

Baseline

- INDi - *Incremental Unsupervised Name Disambiguation* [Carvalho et al., 2011]
- Desambigua apenas os novos registros inseridos no repositório.
- Prioriza a identificação estritamente correta do grupo ao qual o novo registro pertence.
- Pré-processamento dos registros e seleção dos possíveis grupos de autores do registro a ser inserido.

- Processo de desambiguação é dividido em três etapas:
 - ① Verifica se existe ao menos um co-autor em comum, e se há similaridade do título do trabalho ou do título do veículo de publicação entre o novo registro e o grupo selecionado.
 - ② Verifica se o novo registro não possui co-autor, e avalia a similaridade entre título do trabalho ou título do veículo de publicação, com seus respectivos valores parâmetros de similaridade acrescidos pelo incremento.
 - ③ Idêntica a etapa 2, a única diferença é que se verifica a lista de co-autores do grupo selecionado está vazia.
- Caso insucesso nas 3 etapas, o registro é considerado pertencente a um novo autor.

Método Proposto

- Reduzir a quantidade de registros usados no processo de desambiguação.
- Selecionar os registros que melhor representam o grupo ao qual eles pertencem.
- Utilizar somente os registros selecionados no processo de desambiguação.
- Número máximo de registros representativos é igual metade de registros do grupo mais um.

Escolha de Registro Representativo

Algorithm 1 Escolha de registro representativo

Input: Registro *reg*; Grupo *grupo*; Parâmetros de similaridade $\alpha_{coAutores}$, α_{titulo} e $\alpha_{veiculo}$;
Número máximo de registros representativos;

Output: *verdadeiro* se a citação *cit* for representativa e *falso* caso contrário;

```

1 repre ← falso;
2 grupoRepre ← GetGrupoRepresentativo(grupo);
3 if grupoRepre está vazio then
4   return verdadeiro;
5 if ¬CoautoresSimilar(reg, grupoRepre,  $\alpha_{coAutores}$ ) then
6   repre ← verdadeiro;
7 else
8   if ¬TituloSimilar(reg, grupoRepre,  $\alpha_{titulo}$ ) ou ¬VeiculoSimilar(reg, grupoRepre,  $\alpha_{veiculo}$ )
      then
9     repre ← verdadeiro;
10 if repre é verdadeiro then
11   if NumCits(grupoRepre) < maxCit then
12     return verdadeiro;
13   else
14     return removeReg(reg, grupo);
15 return repre;
```

Remoção de Registro Representativo

Algorithm 2 Remove registro representativo

Input: Registro *reg*; Grupo *grupo*; Parâmetros de similaridade α_{autor} , $\alpha_{coAutores}$;

Output: **verdadeiro** se o registro *reg* representar melhor o grupo do que algum registro pertencente ao grupo representativo e **falso** caso contrário;

```
16 remove  $\leftarrow$  falso;
17 grupoRepre  $\leftarrow$  GetGrupoRepresentativo(grupo);
18 for  $r \in$  grupoRepre do
19   if AutorSimilar(reg,r, $\alpha_{autor}$ ) then
20     if CoautoresSimilar(reg,r, $\alpha_{coAutores}$ ) and NumeroCoAutores(reg,r) then
21       removeReg(r);
22       return verdadeiro;
23   else
24     if CoautoresSimilar(reg,r, $\alpha_{coAutores}$ ) and NumTermosTitulo(reg,r) then
25       removeReg(r);
26       return verdadeiro;
27 return remove;
```

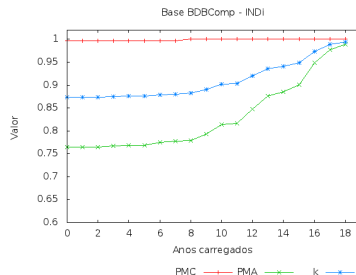
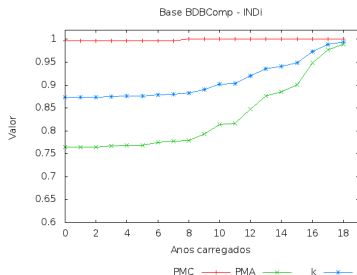
Avaliação Experimental

- Foram utilizadas duas coleções para os experimentos: BDBComp possui 361 registros e KISTI possui 41672 registros.
- Duas situações de experimentos:
 - O repositório da DL não possui registros cadastrados.
 - O repositório da DL já possui registros cadastrados, livres de ambiguidade e com os grupos de autores devidamente classificados.

Tabela : Parâmetros dos Métodos

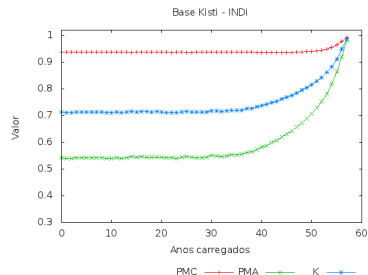
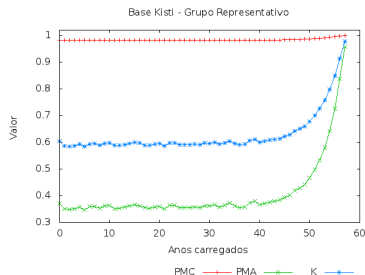
Base	INDi			Grupo Representativo		
	α_{Title}	α_{Venue}	δ	$\alpha_{coAuthors}$	α_{title}	α_{venue}
BDBComp	0.0	0.2	0.2	1	0.0	0.4
Kisti	0.0	0.0	0.2	1	0.0	0.4

Análise dos Experimentos - Coleção BDBComp



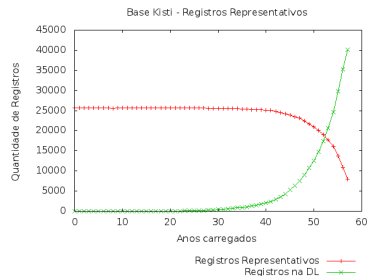
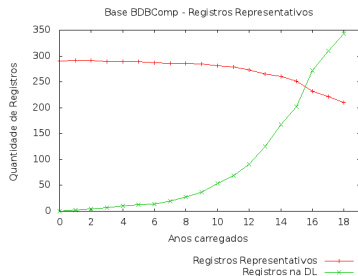
No pior caso, usando 73.41% dos registros como representativo obteve-se um resultado apenas 3.95% inferior ao uso de todos os registros.

Análise dos Experimentos - Coleção KISTI



No pior caso, usando 58.18% dos registros como representativo obteve-se um resultado apenas 19.32% inferior ao uso de todos os registros.

Análise dos Experimentos - Quantidade de Registros Representativos



A porcentagem de registros representativos diminuem de acordo com o aumento dos anos carregados no repositórios das DLs. No entanto, essa queda na coleção BDBComp é pequena.

Conclusão

- Selecionar um conjunto de registros representativos de um grupo.
- Redução do uso de registros no processo de desambiguação com uma pequena perda de eficiência.
- Trabalhos Futuros: pretende-se investigar outras maneiras de selecionar registros representativos, aplicá-los a diversos métodos, bem como realizar experimentos com outras bases.

- A. P. Carvalho, A. A. Ferreira, A. H. F. Laender, and M. A. Gonçalves. Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management*, 2(3):289–304, 2011.
- J. C. French, A. L. Powell, and E. Schulman. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science and Technology*, 51(8):774–786, May 2000. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(2000)51:8<774::AID-ASI90>3.3.CO;2-G. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(2000\)51:8<774::AID-ASI90>3.3.CO;2-G](http://dx.doi.org/10.1002/(SICI)1097-4571(2000)51:8<774::AID-ASI90>3.3.CO;2-G).
- M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transactions on Information Systems*, 22(2):270–312, 2004.
- I. Lapidot. Self-organizing-maps with bic for speaker clustering. *IDIAP Research Report 02-60*, IDIAP Research Institute, 2002.

- D. Lee, B.-W. On, J. Kang, and S. Park. Effective and scalable solutions for mixed and split citation problems in digital libraries. In *Proceedings of the 2nd international workshop on Information quality in information systems*, IQIS '05, pages 69–76, New York, NY, USA, 2005. ACM. ISBN 1-59593-160-0. doi: 10.1145/1077501.1077514. URL <http://doi.acm.org/10.1145/1077501.1077514>.
- V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, 1966.