Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries

Adriano Veloso Anderson A. Ferreira
Marcos André Gonçalves Alberto H.F. Laender

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Proceedings of the 10th annual joint conference on Digital libraries, 2010

Índice

Introdução

- DLs
  - DBLP
  - CiteSeer
  - MEDLINE
  - BDBComp
- Bibliographic Citation Record
  - **Ferreira, A.A. and Veloso, A. and Gonçalves, M.A. and Laender, A.H.F**. Effective self-training author name disambiguation in scholarly digital libraries. JCDL, 2010.

Problema

Ambiguidade de nomes

c_1 A. Ferreira, A. Veloso, M. Gonçalves, and A. Laender.

c_2 A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. Laender.

c_3 A. A. Ferreira, T. B. Ludermir, R. R. B. de Aquino, M. S. Lira and O. N. Neto.

Objetivo

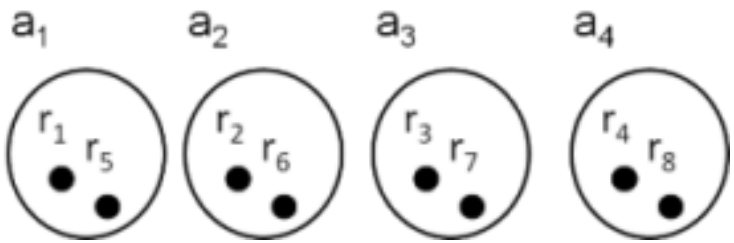
- Função de desambiguação
  - $C = \{c_1, c_2, \ldots, c_k\}$
  - $\{r_1, r_2, \ldots, r_m\}$
  - $\{a_1, a_2, \ldots, a_n\}$
    - a_i contém todos os registros de autoria que representam um mesmo autor

Exemplo

c_1 (r_1) A. Ferreira, (r_2) A. Veloso, (r_3) M. Gonçalves, and (r_4) A. Laender.

c_2 (r_5) A. A. Ferreira, (r_6) A. Veloso, (r_7) M. A. Gonçalves, and (r_8) A. H. Laender.

c_3 (r_9) A. A. Ferreira, (r_{10}) T. B. Ludermir, (r_{11}) R. R. B. de Aquino, (r_{12}) M. S. Lira and (r_{13}) O. N. Neto.



Método Proposto

- Pré-processamento
- Híbrido
  - Não supervisionado
  - Supervisionado

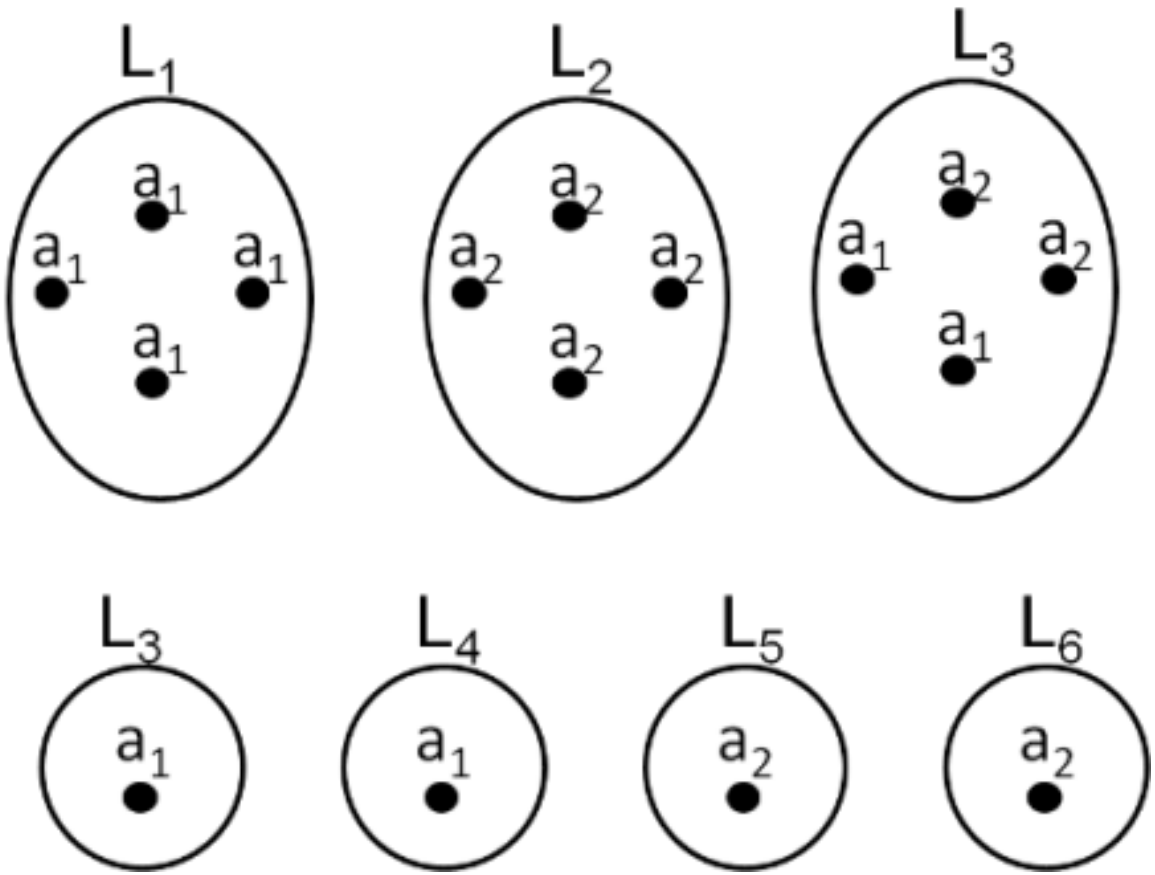
Pré-processamento

- Remoção de stop-words
- Stemming
- Método de bloqueio
  - Grupos ambíguos

Passo não supervisionado

- Entrada
  - registros de autoria (grupos ambíguos)
- Saída
  - exemplos de treinamento
- Tarefas
  - Agrupar registros de autoria em clusters
  - Extrair clusters puros
  - Remover clusters fragmentados

Clustering



Extrair clusters puros

- Para 2 registros de autoria
  - Verifica se ambos os registros tem pelo menos um coautor em comum
  - Então os dois registros são colocados em um mesmo cluster

Remover clusters fragmentados

1. Ordenar os clusters em uma lista C
   - por ordem decrescente de tamanho
2. O maior é inserido no conjunto D (dados de treinamento)
3. O próximo cluster de C a ser inserido, deve ser diferente do que já foi inserido em D
   - Função de similaridade do cosseno
4. O processo termina quando não há mais cluster candidatos a serem inseridos em D
5. Os clusters que restaram são inseridos em T (dados de teste)

Passo supervisionado

- Entrada: Exemplos de treinamento (D)
- Saída: autores corretos em T
- Tarefas
  - Gera função de desambiguação por meio de regras de associação
  - Prevê os autores corretos para os registros de autoria no conjunto de teste T.

Cost-effective on-demand associative author name disambiguation.

Avaliação

- Coleções:
  - DBLP
  - BDBComp
- Métricas:
  - K
  - pF1
- Baseline
  - Métodos não supervisionados: KWAY, SVM-DBSCAN
  - Métodos supervisionados: S-SVM, S-NB

Resultados

DBLP

Método	K	pF1
SAND	**0.712**	**0.630**
KWAY	0.560	0.402
SVM-DBSCAN	0.460	0.279
S-SVM	0.592	0.444
S-NB	0.591	0.449

BDBComp

Método	K	pF1
SAND	**0.842**	**0.597**
KWAY	0.805	0.436
SVM-DBSCAN	0.339	0.088
S-SVM	0.825	0.479
S-NB	0.820	0.457

Referências

Veloso, A. and Ferreira, A.A. and Gonçalves, M.A. and Laender, A.H.F. and Meira, W.. Cost-effective on-demand associative author name disambiguation. Elsevier, 2011.