Cost-effective on-demand associative author name disambiguation

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

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Information Processing & Management, 2011

Índice

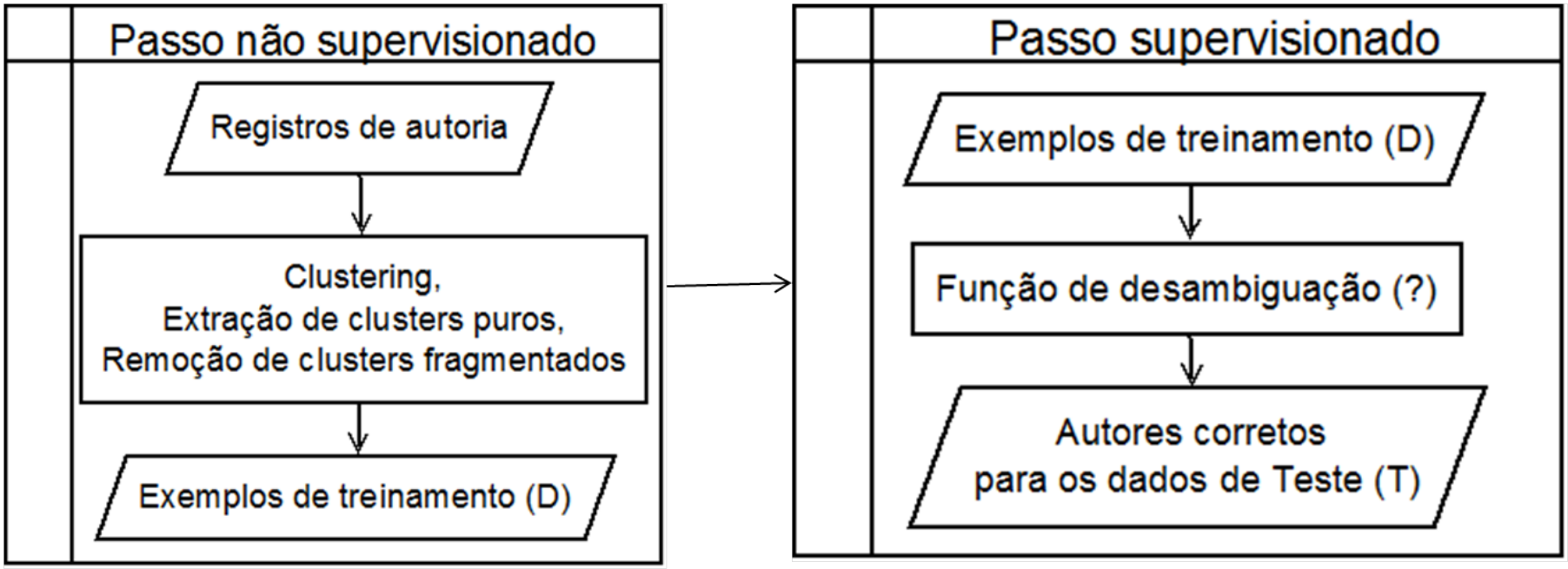
Introdução

- Citação

Adriano Veloso, Anderson A. Ferreira, Marcos André Gonçalves, Alberto H. F. Laender, Wagner Meira Jr. Cost-effective on-demand associative author name disambiguation. Information Processing & Management, 2011.

- $\{f_1, f_2, ..., f_m\}$
  - {coauthor = Anderson A. Ferreira, Marcos André Gonçalves, Alberto H. F. Laender, Wagner Meira Jr}
  - {title = Cost-effective on-demand associative author name disambiguation. }
  - {venue = Information Processing & Management}
  - {year = 2011}

Problema


Passo não supervisionado
Registros de autoria
Clustering,
Extração de clusters puros,
Remoção de clusters fragmentados
Exemplos de treinamento (D)
Passo supervisionado
Exemplos de treinamento (D)
Função de desambiguação (?)
Autores corretos
para os dados de Teste (T)

Desambiguação associativa

- Função de desambiguação
  - Função de desambiguação
    - $\{f_1, f_2, ..., f_m\} \rightarrow \{a_1, a_2, ..., a_n\}$
  - Regras de associação
    - $X \rightarrow a_i$
    - $X \subseteq \{f_1, f_2, ..., f_m\}$
  - Exemplo
    - $\{\text{coauthor} = \text{K. Talwar}, \text{title} = \text{Metric}, \text{venue} = \text{LATIN}\} \rightarrow a_1$

Desambiguação associativa

- $R_{a_i}^x \subseteq R_{a_i} \subseteq R$
  - R é o conjunto de regras arbitrárias
  - R_{a_i}é o conjunto de regras da forma $X \rightarrow a_i$
  - $R_{a_i}^x$é o conjunto de regras da forma $X \rightarrow a_i$ para a citação x
- $\theta = (X \rightarrow a_i)$ mede a força de cada associação
  - coauthor = K. Talwar $\xrightarrow{\theta=1.00} = a_1$
  - venue = LATIN $\xrightarrow{\theta=0.50} = a_1$

Desambiguação associativa

- $s(a_i, x) = \frac{\sum_{j=1}^{|R_{a_i}^x|} \theta(r_j)}{|R_{a_i}^x|}$ medida de confiança média
  - Exemplo
    - coauthor = K. Talwar $\xrightarrow{\theta=1.00} = a_1$
    - venue = LATIN $\xrightarrow{\theta=0.50} = a_1$
    - $s(a_1, c) = \frac{1.00+0.50}{2} = 0.75$
- $\hat{p}(a_i|x) = \frac{s(a_i,x)}{\sum_{j=1}^{|n|} s(a_j,x)}$ normalização de $s(a_i, x)$
  - Exemplo
    - coauthor = K. Talwar $\xrightarrow{\theta=1.00} = a_1$
    - venue = LATIN $\xrightarrow{\theta=0.50} = a_1$
    - $s(a_1, c) = 0.75$
    - $\hat{p}(a_1|c) = \frac{0.75}{0.75} = 1$

Métodos

- 3 desambiguadores baseados em associação:
  - EAND (*Eager Associative Name Disambiguation*)
  - LAND (*Lazy Associative Name Disambiguation*)
  - SLAND (*Self-training LAND*)

EAND

Algorithm 1. Eager Associative Name Disambiguation

Require: Examples in D; σ_{min} , and citation x ∈ T
Ensure: The predicted author of citation x

1: $\pi_{min} \Leftarrow (\sigma_{min} \times |D|)$
2: $R \Leftarrow$ rules r extracted from $D|\pi(r) \geq \pi_{min}$
3: **for each** author a_i **do**
4: $R_{a_i}^{x} \Leftarrow$ rules $X \rightarrow a_i \in R|\pi\ (X \rightarrow a_i) \geq \pi_{min}$ and $X \subseteq x$
5: Estimate $\hat{p}(a_i|x)$
6: **end for**
7: Predict author a_i such that $\hat{p}(a_i|c) > \hat{p}\left(a_j|c\right)\forall j \neq i$

Exemplo

	Label	Coauthors	Publication title	Venue
c_1	a_1	K.Talwar	Doubling Metric	LATIN
c_2	a_1	T. Chan, K. Talwar	Dimensional Embeddings	SODA
c_3	a_1	T. Chan	Approximating TSP	SODA
c_4	a_1	T. Chan (among others)	Metric Embeddings	FOCS
c_5	a_2	T. Ashwin, S. Ghosal	Adaptable Similarity Search	VLDB
c_6	a_2	---	Explanation-Based Failure Recovery	AAAI
c_7	a_2	M. Bhide (among others)	Dynamic Access Control Framework	ICDE
c_8	a_3	S. Sarawagi	Creating Probabilistic DBs	VLDB
c_9	a_3	S. Puradkar (a. others)	Semantic Web Based Pervasive	AAAI
c_{10}	a_4	V. Harinarayan	Virtual Database Technology	ICDE
c_{11}	a_1?	K. Talwar	Approximating Unique Games	SODA
c_{12}	a_4 ?	V. Harinarayan	Index Selection for OLAP	ICDE
c_{13}	a_4 ?	I. Mumick	What is the DW Problem?	VLDB
c_{14}	a_4 ?	V. Harinarayan	Aggregate-Query Processing	VLDB
c_{15}	a_5 ?	J. Hennessy (a. others)	Flexible Use of Memory	ISCA

EAND

- Define
  - $D = 10$ conjunto de treinamento
  - $\sigma_{min} = 0.20$ limite que separa regras frequentes
  - $\pi_{min} \Leftarrow (\sigma_{min} \times |D|) = (0.20 \times 10) = 2$ função de popularidade do autor

c_1	a_1	K. Talwar	Doubling Metric	LATIN
c_2	a_1	K. Talwar	Dimensional Embeddings	SODA
c_{11}		K. Talwar	Approximating Unique Games	SODA

- Extrai a regra
  - coauthor = K. Talwar $\wedge$ venue = LATIN $\xrightarrow{\theta=1.00} = a_1$
  - $\hat{p}(a_i|c_{11}) = 1.00$
  - Portanto preve que o autor a_1é o autor correto para a citação c_{11}

LAND

Algorithm 2. Lazy Associative Name Disambiguation

Require: Examples in D; σ_{min} , and citation x ∈ T
Ensure: The predicted author of citation x

1: Let $\mathcal{L}(f_i)$ be the set of examples in D in which feature f_i has occurred
2: $D^x \Leftarrow \emptyset$
3: **for each** feature $f_i \in x$ **do**
4: $\quad D^x \Leftarrow D^x \cup \mathcal{L}(f_i)$
5: **end for**
6: $\pi^x_{min} \Leftarrow (\sigma_{min} \times |D^x|)$
7: **for each** author a_i **do**
8: $\quad R^x_{a_i} \Leftarrow$ rules $X \rightarrow a_i$ extracted from $D^x | \pi\ (X \rightarrow a_i) \geq \pi^x_{min}$
9: $\quad$ Estimate $\hat{p}(a_i|x)$
10: $\quad$ **end for**
11: Predict author a_i such that $\hat{p}(a_i|c) > \hat{p}(a_j|c) \forall j \neq i$

Exemplo

	Label	Coauthors	Publication title	Venue
c_1	a_1	K.Talwar	Doubling Metric	LATIN
c_2	a_1	T. Chan, K. Talwar	Dimensional Embeddings	SODA
c_3	a_1	T. Chan	Approximating TSP	SODA
c_4	a_1	T. Chan (among others)	Metric Embeddings	FOCS
c_5	a_2	T. Ashwin, S. Ghosal	Adaptable Similarity Search	VLDB
c_6	a_2	---	Explanation-Based Failure Recovery	AAAI
c_7	a_2	M. Bhide (among others)	Dynamic Access Control Framework	ICDE
c_8	a_3	S. Sarawagi	Creating Probabilistic DBs	VLDB
c_9	a_3	S. Puradkar (a. others)	Semantic Web Based Pervasive	AAAI
c_{10}	a_4	V. Harinarayan	Virtual Database Technology	ICDE
c_{11}	a_1?	K. Talwar	Approximating Unique Games	SODA
c_{12}	a_4 ?	V. Harinarayan	Index Selection for OLAP	ICDE
c_{13}	a_4 ?	I. Mumick	What is the DW Problem?	VLDB
c_{14}	a_4 ?	V. Harinarayan	Aggregate-Query Processing	VLDB
c_{15}	a_5 ?	J. Hennessy (a. others)	Flexible Use of Memory	ISCA

LAND

- D^x é o novo conjunto de treinamento composto por todas as citações que correspondem com a citação x no conjunto de teste.

D^c_{12}	Label	Coauthors	Publication title	Venue
c_7	a_2	---	----	ICDE
c_{10}	a_4	V. Harinarayan	----	ICDE

- Regras de associação para a citação c_{12}:

$$\text{coauthor} = \text{V. Haribarayan} \xrightarrow{\theta=1.00} = a_4$$
$$\text{coauthor} = \text{V. Haribarayan} \wedge \text{venue} = \text{ICDE} \xrightarrow{\theta=1.00} = a_4$$
$$\text{venue} = \text{ICDE} \xrightarrow{\theta=0.50} = a_4$$
$$\text{venue} = \text{ICDE} \xrightarrow{\theta=0.50} = a_2$$

LAND

$$s(a_4, c_{12}) = \frac{1.00+1.00+0.50}{3} = 0.83$$
$$\hat{p}(a_4|c_{12}) = \frac{0.83}{0.50+0.83} = 0.62$$
$$s(a_2, c_{12}) = \frac{0.50}{1} = 0.50$$
$$\hat{p}(a_2|c_{12}) = \frac{0.5}{0.50+0.83} = 0.38$$

- Portanto, a_4 é o autor correto para a citação c_{12}, pois $\hat{p}(a_4|c_{12})$é maior do que $\hat{p}(a_2|c_{12})$

SLAND

- Expansão de LAND
- Tenta resolver os seguintes problemas:
  - Quando a probabilidade de dois autores é a mesma
  - Quando o conjunto $D^x = \emptyset$

SLAND

Algorithm 3. Self-training LAND

Require: Examples in D; σ_{min} , Δ_{min} , γ_{min} and citation x ∈ T
Ensure: The predicted author of citation x (if the prediction is not abstained)
(The ten first steps are exactly the same ones shown in Algorithm 2, and thus they are omitted here)
⋮
[11:] **if** $\gamma(x) \geq \gamma_{min}$ **then**
[12:] Create a new label, a_k
[13:] Predict author a_k
[14:] Include $\{x \cup a_k\}$ in D
[15:] **else if if** $\Delta(x) \geq \Delta_{min}$ **then**
[16:] Predict author a_i such that $\hat{p}(a_i|c) > \hat{p}(a_j|c) \forall j \neq i$
[17:] Include $\{x \cup a_i\}$ in D
[18:] **else**
[19:] Place x in the end of the queue
[20:] **end if**

SLAND

Inclui novos exemplos nos dados de treinamento

1. $\Delta(x) = \frac{\hat{p}(a_i|x)}{\sum_{j=1}^{|n|} \hat{p}(a_j|x)}$ medida de confiabilidade
2. Δ_{min} é um parâmetro especificado pelo usuário
   - se $\Delta(x) \geq \Delta_{min}$

Não inclui exemplos que não são confiáveis

- Coloca as citações randomicamente em uma fila de prioridades

1. Olha a citação no início da fila e verifica a sua confiabilidade
   - se $\Delta(x) < \Delta_{min}$
     - É colocada no final da fila
   - se $\Delta(x) \geq \Delta_{min}$
     - inclui como novo exemplo nos dados de treinamento

SLAND

Encontra novos autores

1. $\gamma(x)$ é o número de regras extraídas para D^x
2. γ_{min} é um parâmetro especificado pelo usuário

- se $\gamma(x) < \gamma_{min}$
  - cria um novo rótulo a_k
  - inclui como novo exemplo nos dados de treinamento

Exemplo

c_{13}	a_4 ?	I. Mumick	What is the DW Problem?	VLDB
c_{14}	a_4 ?	V. Harinarayan	Aggregate-Query Processing	VLDB
c_{15}	a_5 ?	J. Hennessy (a. others)	Flexible Use of Memory	ISCA

- $\Delta_{min} = 1.50$

D^c_{13}	Label	Coauthors	Publication title	Venue
c_5	a_2	---	----	**VLDB**
c_8	a_3	---	----	**VLDB**

Exemplo

- Extrai as regras:

$$\text{venue} = \text{VLDB} \xrightarrow{\theta=0.50} = a_4$$
$$\text{venue} = \text{VLDB} \xrightarrow{\theta=0.50} = a_2$$

1. $\Delta(c_{13}) = \frac{\hat{p}(a_2|c_{13})}{\hat{p}(a_3|c_{13})} = \frac{0.50}{0.50} = 1 < \Delta_{min}$
2. c_{13} é colocada no final da fila

Exemplo

D^c_{14}	Label	Coauthors	Publication title	Venue
c_8	a_3	---	----	**VLDB**
c_{10}	a_4	**V. Harinarayan**	----	**ICDE**

1. $\Delta(c_{13}) = \frac{\hat{p}(a_4|c_{13})}{\hat{p}(a_3|c_{13})} = \frac{0.75}{0.50} = 1.50 \geq \Delta_{min}$
2. c_{14} é incluída nos dados de treinamento e o autor correto é a_4, pois $\hat{p}(a_4|c_{13})$ é maior que $\hat{p}(a_3|c_{13})$.

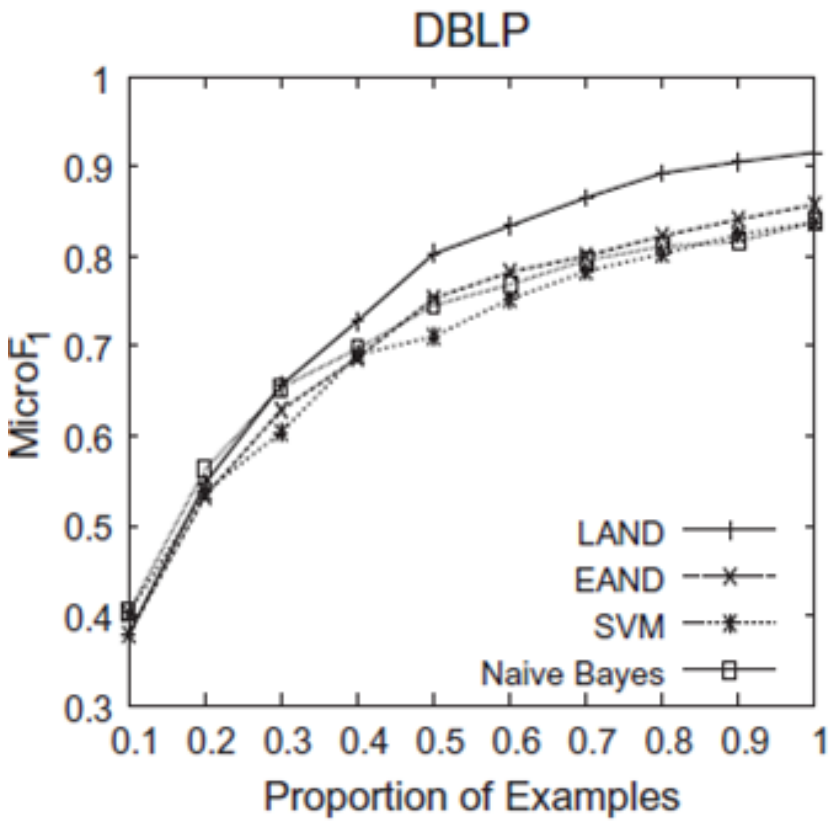
Exemplo

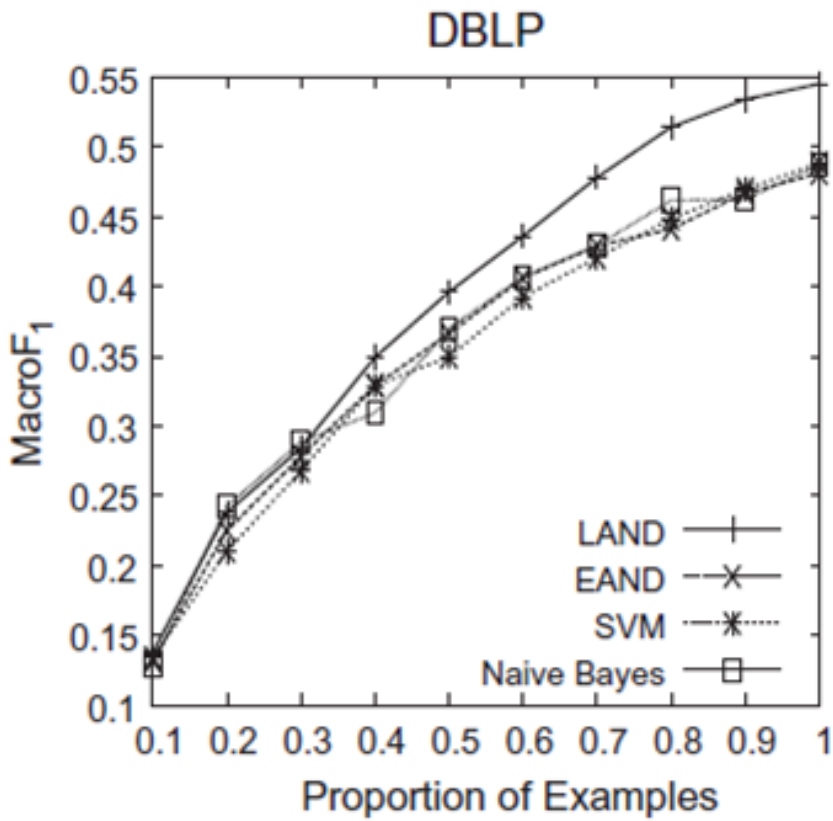
1. $D^{c_{15}} = \emptyset$
   1. $\gamma_{min} = 1$
   2. $\gamma(c_{15}) = 0$ (o número de regras extraídas é igual a zero)
2. é criado um novo autor a_{15} para c_{15} e é incluída nos dados de treinamento

Avaliação

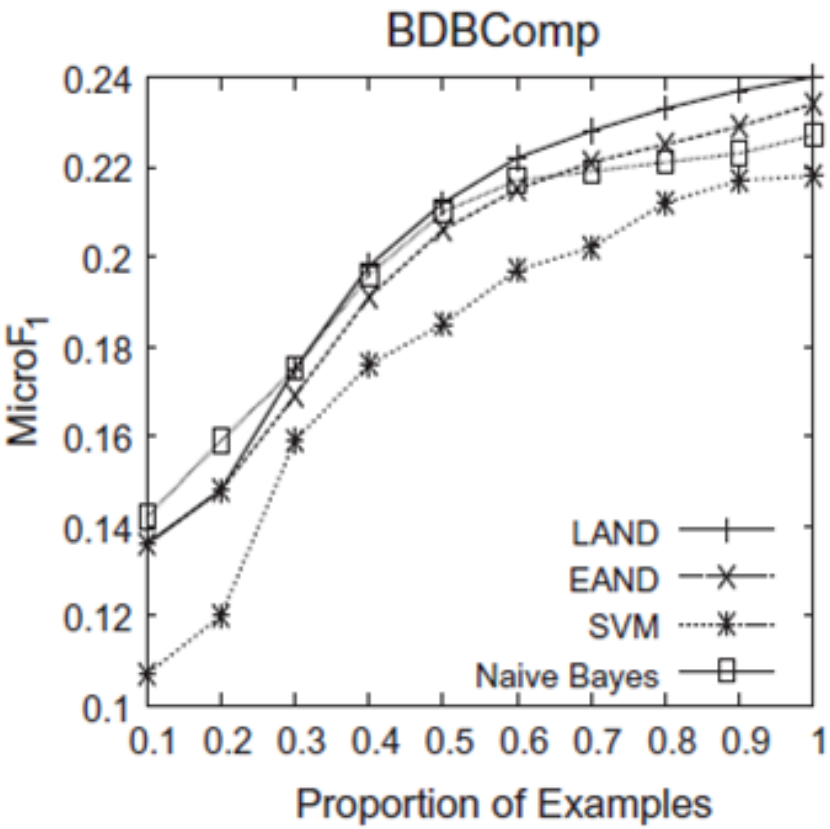
- Coleções:
  - DBLP
  - BDBComp
- Métricas:
  - F1
    - microF1 = é a média entre autores específicos e sobre todos os autores.
    - macroF1= decisões para todos os autores foram contadas em um conjunto comum
- Baseline
  - Métodos supervisionados: S-SVM, S-NB

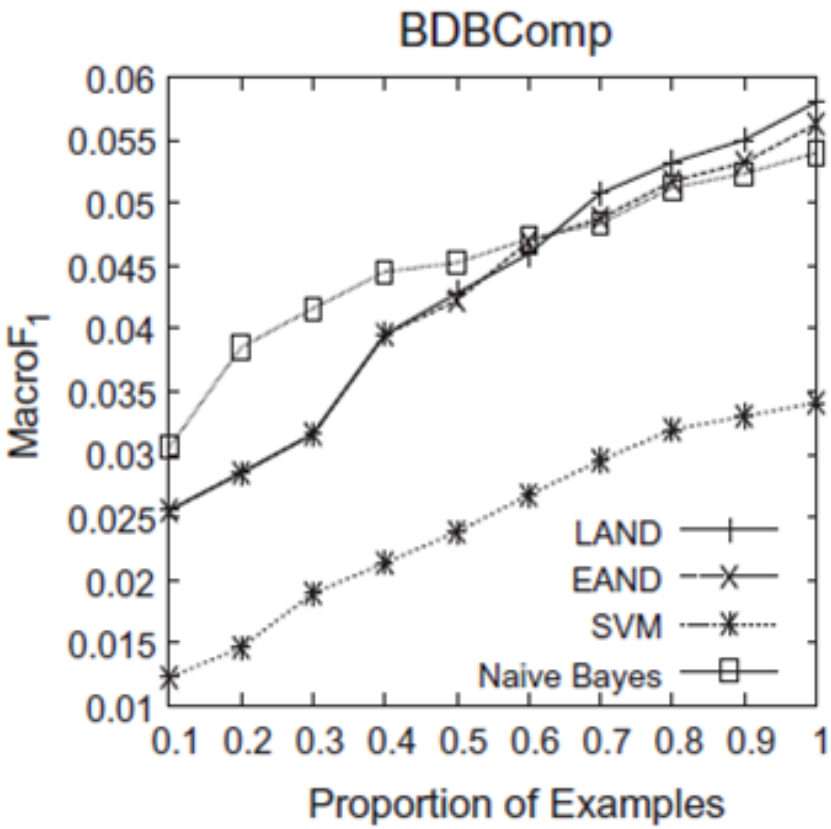
Resultados


DBLP
MicroF1
Proportion of Examples
LAND
EAND
SVM
Naive Bayes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


DBLP
MacroF1
Proportion of Examples
LAND
EAND
SVM
Naive Bayes
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Resultados


BDBComp
MicroF1
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Proportion of Examples
LAND
EAND
SVM
Naive Bayes


BDBComp
MacroF1
0.06
0.055
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Proportion of Examples
LAND
EAND
SVM
Naive Bayes