UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Multi-task Active Learning in Entity Resolution

Henrique Saccani

Relatore: Prof. Giovanni Simonini

2021

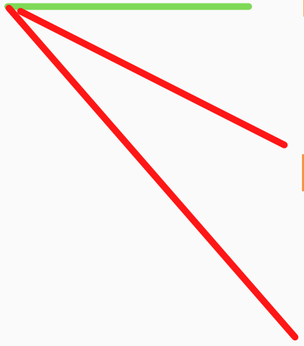
Entity resolution : Matching

A

	Title	Director	Country	Cast
A1	Battleship Potemkin	Sergei Eisenstein	Soviet Union	Aleksandr Antonov Vladimir Barksy Grigori Aleksandrov
A2	The Hateful Eight	Quentin Tarantino	US	Samuel L. Jackson, Kurt Russell,
	...			
A3	Frozen	Adam Green	US	Emma Bell, Shawn Ashmore, Kevin Zegers
	...			

Rogerbert.com

3k records



B

	Title	Director	Country	Cast
B1	The Battleship Potemkin	S. Eisenstein	USSR	A. Antonov V. Barksy G. Aleksandrov
B2	The ***H8ful*** Eight	Q. Tarantino	USA	K. Russell, W. Goggins, S. L. Jackson, J. Jason Leigh
	...			
B3	Frozen	C. Buck	USA	K. Bell, I. Menzel, J. Groff
	...			

Imdb.com

3k records

- *Identificano la stessa entità nel mondo reale (Match)*
- *Non Identificano la stessa entità nel mondo reale (non Match)*

Entity resolution : Matching

Quantificare la somiglianza della coppia

A1

Title	Director	Country	Cast
Battleship Potemkin	Sergei Eisenstein	Soviet Union	Aleksandr Antonov Vladimir Barksy Grigori Aleksandrov

B1

Title	Director	Country	Cast
The Battleship Potemkin	S. Eisenstein	USSR	A. Antonov V. Barksy G. Aleksandrov

A1 e B1 sono match

id1	id2	Js(title)	Lev(title)	...	Js(cast)	Lev(cast)	Label
A1	B1	0.82	0.76	...	0.57	0.64	1
A1	B2	0.21	0.29	...	0.3	0.1	0
A3	B3	1	1	...		0.29	0

Attraverso i risultati delle similarità il classificatore è capace di fare previsioni

1 = match

0 = non match

Entity resolution : Matching

Come identificare se sono match?

A1

Title	Director	Country	Cast
Battleship Potemkin	Sergei Eisenstein	Soviet Union	Aleksandr Antonov Vladimir Barksy Grigori Aleksandrov

B2

Title	Director	Country	Cast
The ***H8ful*** Eight	Q. Tarantino	USA	K. Russell, W. Goggins, S. L. Jackson, J. Jason Leigh

A1 e B2 non sono match

id1	id2	Js(title)	Lev(title)	…	Js(cast)	Lev(cast)	Label
A1	B1	0.82	0.76	…	0.57	0.64	1
A1	B2	0.21	0.29	…	0.3	0.1	0
A3	B3	1	1	…		0.29	0

Attraverso i risultati delle similarità il classificatore è capace di fare previsioni

1 = match

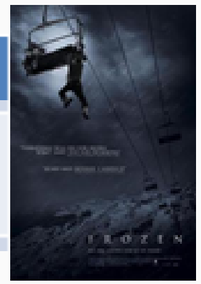
0 = non match

Entity resolution : Matching

Come identificare se sono match?

A3

Title	Director	Country	Cast
Frozen	Adam Green	US	Emma Bell, Shawn Ashmore, Kevin Zegers



B3

Title	Director	Country	Cast
Frozen	C. Buck	USA	K. Bell, I. Menzel, J. Groff

A3 e B3 non sono match (anche se hanno lo stesso nome)

id1	id2	Js(title)	Lev(title)	...	Js(cast)	Lev(cast)	Label
A1	B1	0.82	0.76	...	0.57	0.64	1
A1	B2	0.21	0.29	...	0.3	0.1	0
A3	B3	1	1	...		0.29	0

Attraverso i risultati delle similarità il classificatore è capace di fare previsioni

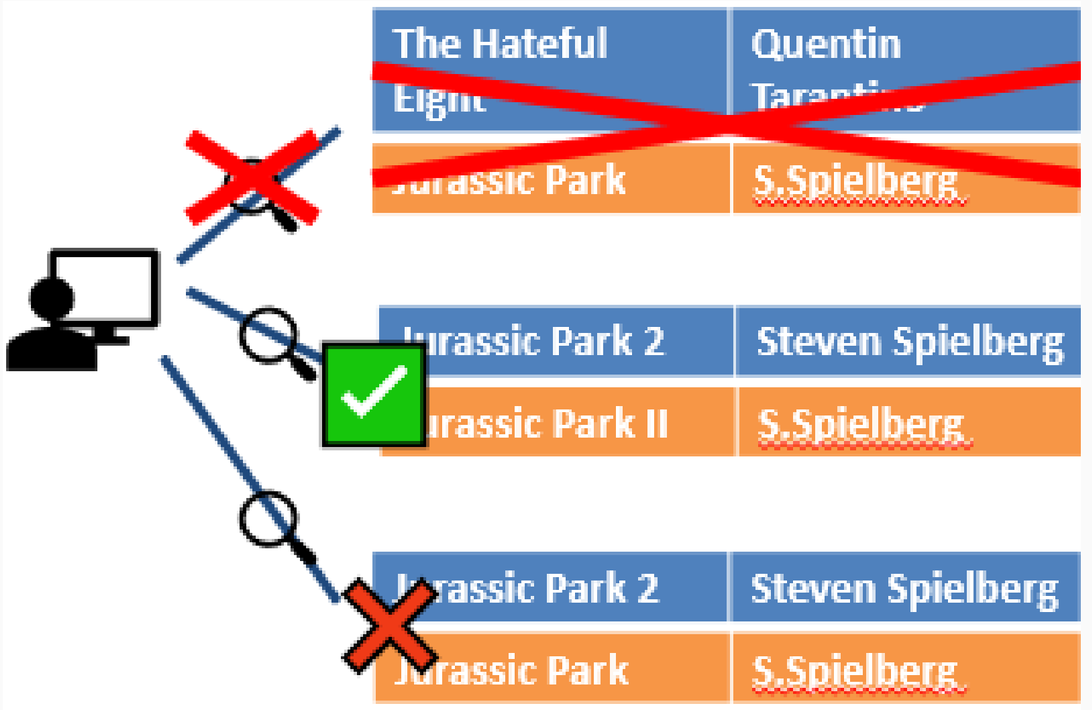
1 = match

0 = non match

Entity resolution : Scalabilità

Come evitare di comparare coppie molto diverse?

- Troppe coppie da comparare:
  ex : 3000 x 3000 = 9 M **(n^2)**
- Comparazioni approfondite richiedono molte risorse
- Coppie molto diverse non necessitano di comparazioni esaustive

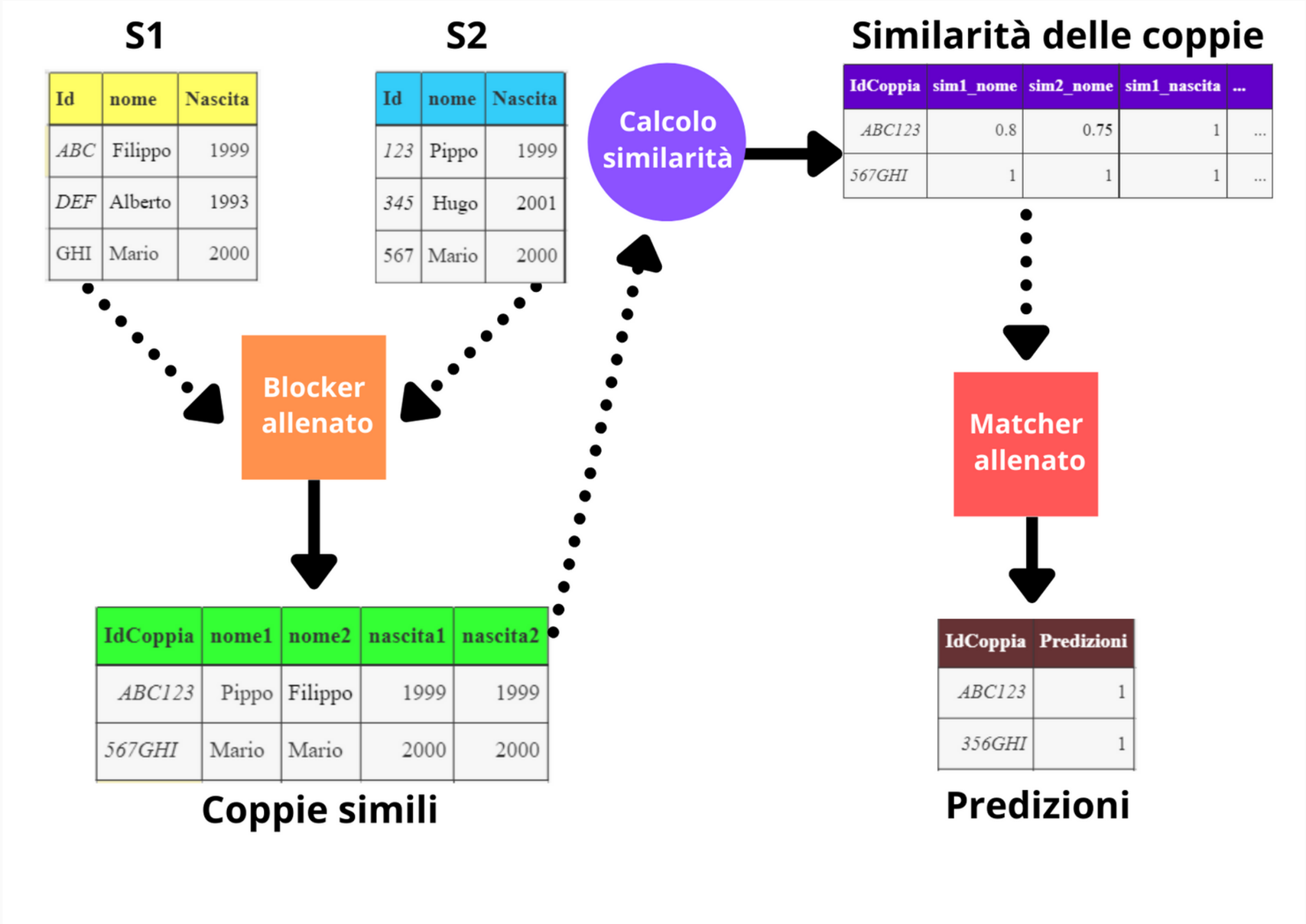


Blocking

Filtrare possibili coppie

- Il blocking utilizza funzioni con meno costo computazionale per attuare come filtro
  - Crea multipli blocchi che rappresentano certe similarità e inserisce i profili in uno o più blocchi
  - Coppie non presenti in almeno un stesso blocco sono scartate
- Utilizzato un modello di **Meta-Blocking**
  - Si utilizza di *Machine Learning* per creare blocchi
  - Impiega un classificatore binario

Flusso di Lavoro



Allenamento iniziale

Come massimizare l'utilita delle coppie etichettate

- Etichettare dati è un processo dispendioso
  - L'utente non ha tempo infinito
  - Molte volte ha costi economici
- Quali coppie etichettare per l'allenamento?
  - Scegliere **coppie simili** per allenare il classificatore porta a risultati migliori
  - Evitare non match ovvi

Active Learning

Quali coppie utilizzare per l'allenamento iniziale?

- Sub campo di Machine Learning
- AL permette al classificatore *scegliere* quali dati vorrebbe fossero utilizzati per il suo allenamento
  - *obbiettivo :* etichettare coppie più utili all'apprendimento del modello
- Strategia *Uncertainty sampling*
  - Identifica i dati dove l'algoritmo ha meno fiducia nella sua previsione
  - Dati di questo tipo permettono al classificatore di imparare piu velocemente

Blocking e Matching : Allenamento

Due classificatori binari per un unico problema di ER

- I dati labellati utilizzati per allenare il blocker possono allenare anche il matcher
  - Possibilità di labellare meno dati
  - Spendere meno risorse

id1	id2	ND	...	CFIBF	Label
A1	B1	0.65	...	0.43	1
A1	B2	0.83	...	0.37	0
A3	B3	0.95	...		0

id1	id2	Js(title)	Lev(title)	...	Js(cast)	Lev(cast)	Label
A1	B1	0.82	0.76	...	0.57	0.64	1
A1	B2	0.21	0.29	...	0.3	0.1	0
A3	B3	1	1	...		0.29	0

Blocker ***Matcher***

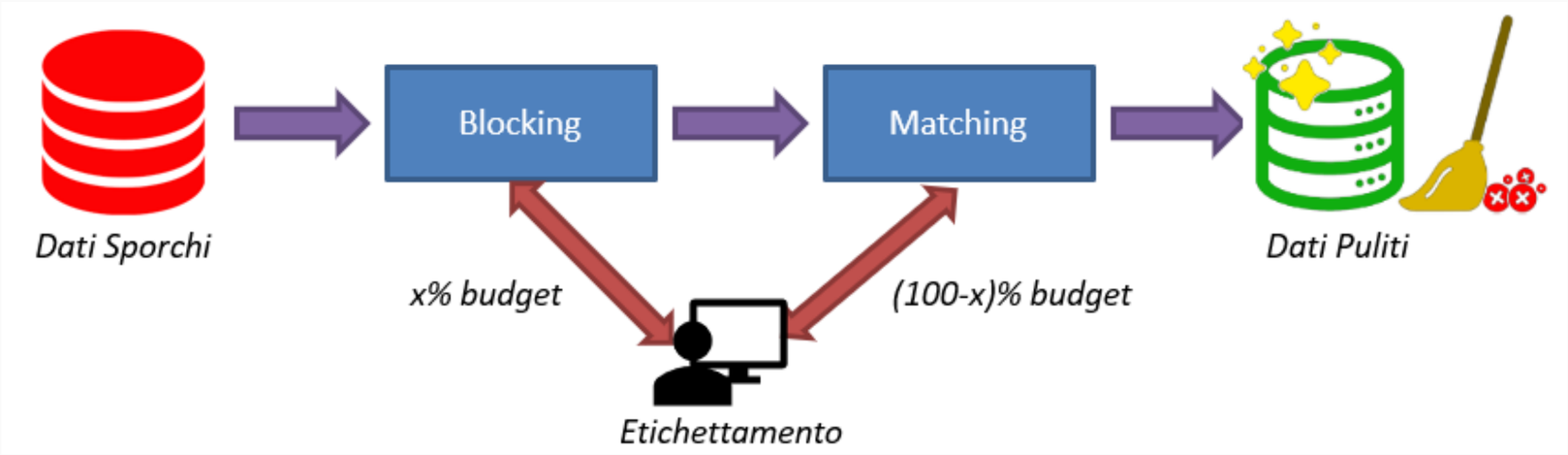
Match

Non Match

Blocking e Matching : Allenamento

Quantità di dati etichettati utilizzati da ogni learner

- La quantità totale di dati etichettati utilizzati nel problema è chiamata di **Budget**
  - *x% budget* = coppie scelte dal AL del blocker
  - *(100-x)% budget* = coppie scelte dal AL del matcher
- Il blocker è allenato con solo con x% budget
- Il matcher è allenato con **tutte** le coppie del budget



Blocking e Matching : Allenamento

Obbiettivi

- Massimizzare precision e recall
- Identificare trade-offs al variare delle percentuali di budget utilizzate nel blocker e nel matcher
- Comparare il nostro approccio multi task a soluzioni allo stato dell'arte

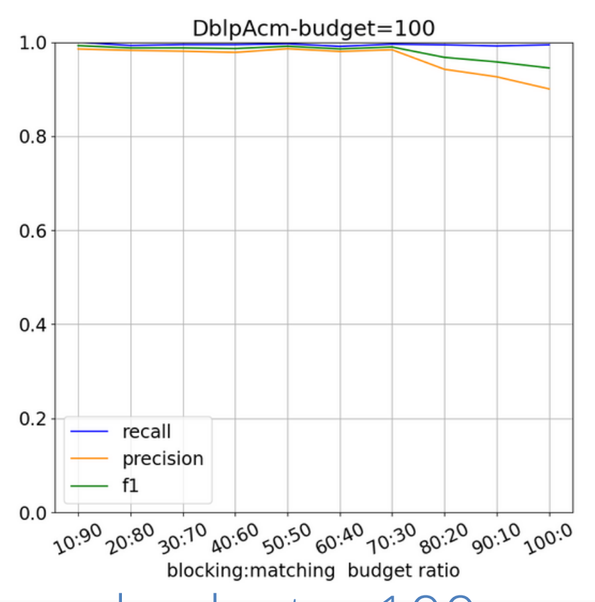
Dettagli implementativi

Nome	subdataset1	subdataset2	matches	tipo
DblpAcm	2.6k	2.3k	2.2k	Strutturato
ScholarDblp	2.5k	61.3k	2.3k	Strutturato
AbtBuy	1.1k	1.1k	1.1k	Sporco
AmazonGoogleProd.	1.4k	3.3k	1.3k	Sporco

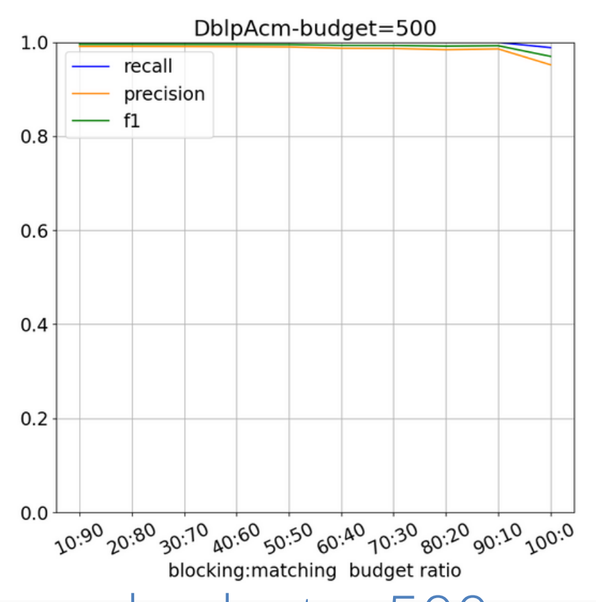
- Esperimenti realizzati in 4 dataset distinti
- **Algoritmi di similarità** diversi a seconda del dataset
- **Budget**
  - Due budget diversi : 100 e 500
  - budget ratio (%blocking : %matching)
  - Esempi di distribuzione di budget : 10:90, 20: 80...100:0
- Classificatori **random forest** sia per il matcher che per il blocker
  - Interpretabilità

Risultati : Dataset strutturati

Recall, Precision e f1-score al variare del budget-ratio

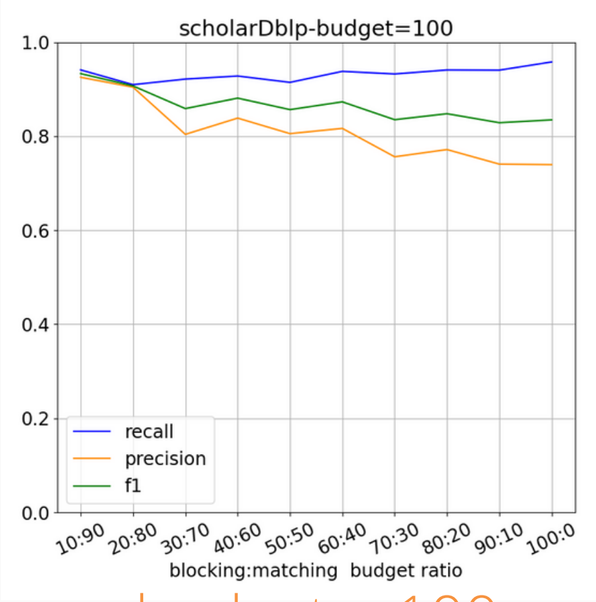


budget = 100

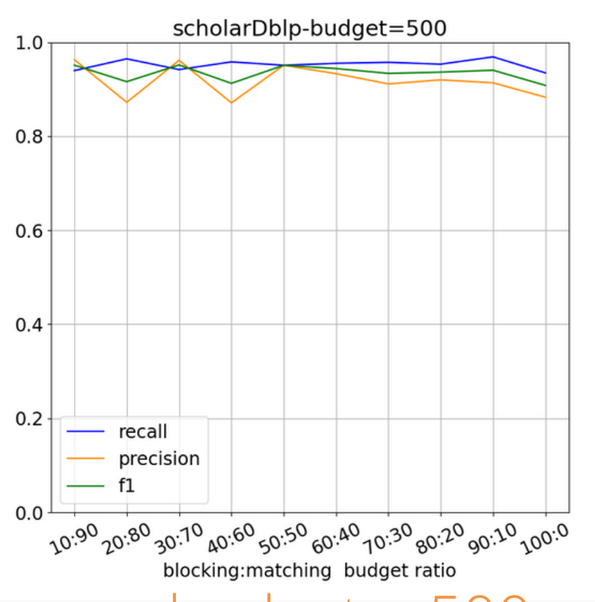


budget = 500

Dataset Dblp-ACM



budget = 100



budget = 500

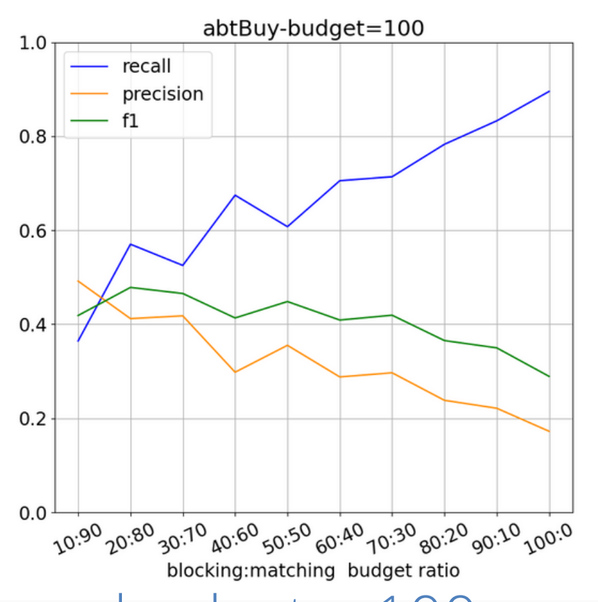
Dataset Dblp-Scholar

10:90 ← **100:0**

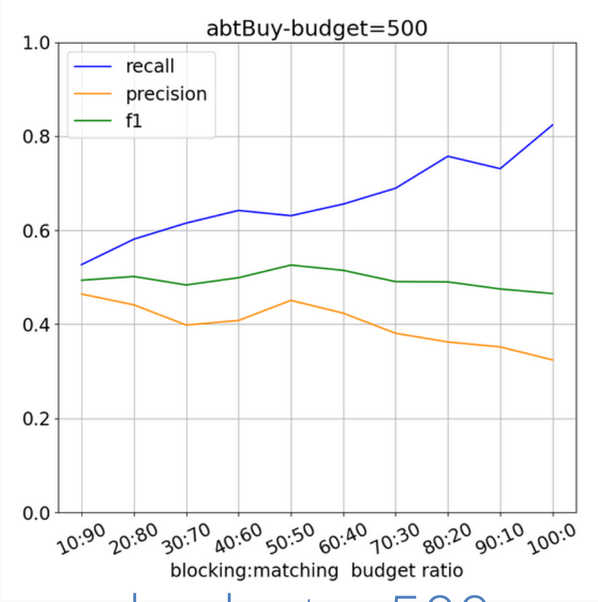
F1 Maggiore all'aumentare del %matching

Risultati : Dataset sporchi

Recall, Precision e f1-score al variare del budget-ratio

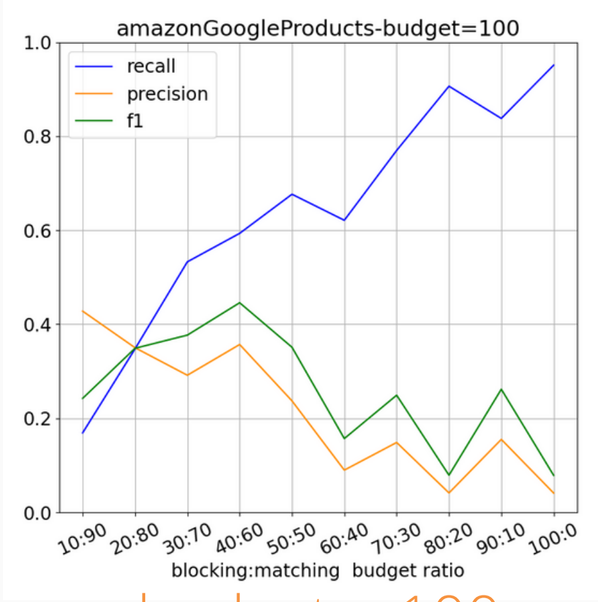


budget = 100

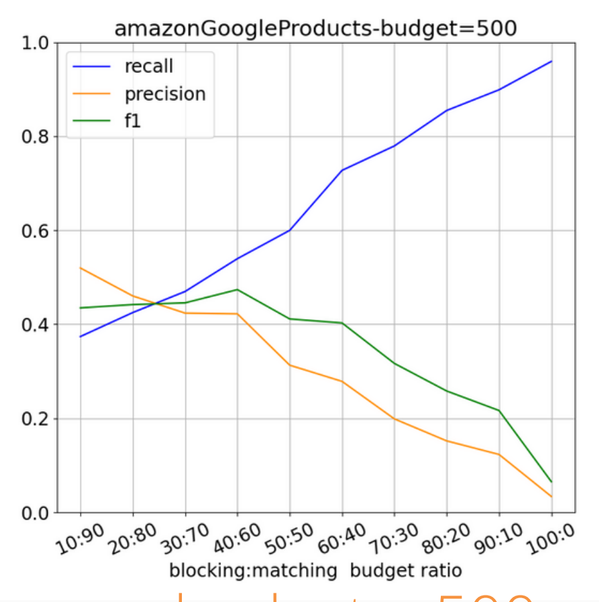


budget = 500

Dataset AbtBuy



budget = 100



budget = 500

Dataset Amazon Google

10:90 ← → **100:0**

Trade-off tra recall e precision al variare dell'allocazione

Risultati

Comparazione con soluzioni allo stato dell'arte

- Una configurazione **40:60** di distribuzione del budget funziona bene in tutti i dataset
- In **dataset strutturati** è stato possibile ottenere f-score prossimi alle soluzioni allo stato dell'arte che utilizzano più di 9k coppie etichettate
- In **dataset sporchi** risultati prossimi alle soluzioni di Magellan che utilizzano più di 6k coppie etichettate

Dataset	Budget ratio	AL-100	AL-500	DeepMatcher	Magellan
abtBuy	40:60	41.3	50.0	62.8	*43.6*
amazon-google	40:60	44.6	47.4	69.3	*49.1*
DBLP-ACM	40:60	98.6	99.5	98.4	98.4
DBLP-Scholar	40:60	88.1	91.2	94.7	92.3

Tabella 3.3: F1-score ottenuti con budget ratio 40:60

Considerazioni Finali

- AL può essere facilmente implementato per raggiungere performance prossime a soluzioni allo stato dell'arte
- Un' allocazione del budget 40:60 (blocking:matching) ha dimostrato di funzionare bene con tipi diversi di dataset
- In lavori futuri sarebbe interessante identificare dinamicamente la configurazione dell' allocazione del budget più performante per ogni dataset

Referenze

G. Simonini, H. Saccani, L. Gagliardelli, L. Zecchini, D. Beneventano and S. Bergamaschi: The Case for Multi-task Active Learning Entity Resolution (2021). In 29th Italian Symposium on Advanced Database Systems.