STUDIO E SPERIMENTAZIONE DI UN ALGORITMO PER L'ENTITY MATCHING BASATO SULLE ESPRESSIONI REGOLARI

Lisa Trigiante

Prof. Sonia Bergamaschi
PhD. Giovanni Simonini
Dott. Luca Zecchini

ENTITY RESOLUTION

Ha come obiettivo l'individuazione all'interno di un insieme di dati (dataset) degli elementi (record) che rappresentano la stessa entità del mondo reale.

ID	MARCA	MODELLO	DIMENSIONE SCHERMO
www.ebay.com//143	HP	e21	17
ce.yikus.com//12	Dell	e173	17
www.ohc24.ch//67	Hewlett Packard	e201	19
www.pc-canada.com//253	HP	e201	NULL
www.ebay.com//1098	Asus	vn247h	19

ENTITY RESOLUTION E BIG DATA

Con l'avvento dei Big data, si complica il problema di Entity Resolution che per sua natura comporta una complessità quadratica.
Per effettuare la risoluzione delle corrispondenze tra le coppie di tuple è necessario che ogni record sia confrontato con tutti gli altri, ad un aumento dei record corrisponde quindi una crescita quadratica dei possibili confronti, che diventa computazionalmente insostenibile nel contesto dei Big Data, dove si hanno milioni di record.

ENTITY RESOLUTION

Fase di blocking

Permette di ridurre il **numero di confronti** dividendo il dataset in **blocchi**, secondo un dato criterio di similarità, effettuando i confronti solo all'interno di ciascun blocco.

Fase di matching

Applica una **funzione di matching** sulle coppie candidate per verificare se si riferiscono o meno alla stessa entità.

Fase di resolution

Dai record che si riferiscono alla stessa entità ottiene un unico **record rappresentativo**, applicando una **funzione di aggregazione** su ogni attributo.

ENTITY RESOLUTION

ID	MARCA	MODELLO	DIMENSIONE SCHERMO
www.ebay.com//143	HP	e21	17
www.ohc24.ch//67	Hewlett Packard	e201	19
www.pc-canada.com//253	HP	e201	NULL

UNION RANDOM VOTE MAX

ID	MARCA	MODELLO	DIMENSIONE SCHERMO
www.ebay.com//143 www.ohc24.ch//67 www.pc-canada.com//253	HP	e201	19

CONTESTO

- Problema: Entity Matching su un dataset di 16k annunci relativi a monitor estratti da 26 siti di e-commerce differenti (DI2KG Programming Contest);
- Obiettivo: massimizzare l'F-score (media armonica di precision e recall) valutato su ground truth di 2.200 specifiche;
- Soluzione: generalizzare un algoritmo ideato dal Dott. Luca Zecchini per un problema analogo nell'ambito dell'ACM SIGMOD Programming Contest, tale algoritmo è stato ampliato al dataset di monitor e sono state gestite particolari eccezioni specifiche dei record contenuti in tale dataset;

ALGORITMO

Il titolo dell'annuncio (page title) è l'unico attributo presente in tutte le specifiche e contiene in genere gli elementi necessari per l'identificazione del monitor:

- MARCA individuata tramite una lista di marchi comuni;
- MODELLO estratto con l'uso di un'espressione regolare poiché in genere costituito da una stringa contenente lettere e numeri;

Un'espressione regolare (regex) è un modo dichiarativo per descrivere un linguaggio regolare.

Le regex sono espressioni formali, composte da una sequenza di simboli, che permettono di rappresentare i linguaggi regolari.

ALGORITMO

La prima fase consiste nell'individuazione di marca e modello mediante varie operazioni:

- Lettura dei singoli record (file JSON);
- Normalizzazione dell'attributo page title (lowercase, punteggiatura e alias);
- Estrazione della marca (lista di brand comuni);
- Gestione di prefissi e suffissi specifici per il marchio;
- Estrazione del modello (espressioni regolari) con gestione dei casi particolari (eccezioni, modelli noti);
- Operazioni aggiuntive sul modello rilevato (equivalenze, rimozione di prefissi e suffissi);
- Distinzione dei record risolti da quelli non risolti (lista solved, lista unsolved);

ALGORITMO

La seconda fase consiste nel confronto di coppie di record per valutare se siano rappresentativi dello stesso oggetto del mondo reale:

- Per i record contenuti nella lista solved vengono confrontati gli attributi marca e modello;
- Per i record contenuti nella lista unsolved viene confrontato l'attributo page title;

Nel caso di una perfetta corrispondenza tra gli attributi confrontati si rileva un caso di match, altrimenti non match.

L'algoritmo produce come risultato un file CSV contenente tutte le coppie di record il cui confronto ha prodotto esito positivo (match).

RISULTATI E CONCLUSIONI

I risultati raggiunti, valutati sul ground truth, sono rispettivamente:

- Precision = 100%
- Recall = 97,2%
- F-score = 98,5%

Il punto di forza dell'algoritmo è la possibilità di adattamento ai casi specifici, che permette di raggiungere valori di precision e recall elevati.

Il punto debole dell'algoritmo consiste nel lavoro umano necessario per tale adattamento (studio manuale dei pattern specifici di ogni singola marca).

GRAZIE PER L'ATTENZIONE