PROGETTI CONCLUSI II CALL

Titolo progetto	MinHash: uno strumento flessibile per integrare basi di dati di grandi dimensioni

Descrizione

L'acquisizione, il trattamento e l'utilizzo a fini statistici di fonti amministrative è alla base del processo di modernizzazione dell'Istituto come testimonia il nuovo modello di Censimento. Poiché nessun archivio è sufficiente, da solo, a soddisfare i requisiti di qualità e di dettaglio informativo che l'Istituto deve garantire, la fruizione del dato amministrativo per fini statistici è legata alla capacità di creare registri integrati le cui unità statistiche provengano da più fonti di dati sia amministrativi che d'indagine. Analogamente al metodo SimHash (Charikar, 2002) già sperimentato con successo in Istituto e disponibile nel toolkit RELAIS (REcord Linkage At IStat) a partire dalla versione 3.0, MinHash è in grado di condensare in modo efficiente il patrimonio informativo contenuto nei dati originari, riducendo drasticamente la complessità dell'operazione di individuazione dei grappoli di record potenzialmente riconducibili ad una medesima unità statistica. Ciò accade perché la ricerca dei possibili duplicati viene effettuata tra "impronte" dei record originari (generalmente stringhe di testo più o meno articolate) che ne preservano le caratteristiche d'interesse ma che sono molto più compatte e semplici da confrontare. La costruzione delle impronte avviene attraverso l'utilizzo ripetuto di funzioni hash (o permutazioni pseudo-casuali) tra loro indipendenti applicate alle stringhe originarie opportunamente scomposte. In particolare, il processo di compressione ha la proprietà di preservare, con una certa approssimazione, la distanza/somiglianza tra i record di partenza: SimHash, ad esempio, approssima la distanza coseno tra le rappresentazioni vettoriali (metodo geometrico) dei record originari, mentre MinHash approssima l'indice di somiglianza di Jaccard tra gli stessi record rappresentati sotto forma di insiemi (metodo combinatorio). L'errore di approssimazione ha una distribuzione normale e diminuisce rapidamente al crescere del numero delle funzioni hash utilizzate, secondo la legge dei grandi numeri (è nullo in valore atteso). La natura probabilistica del processo di compressione consente di confrontare le impronte "localmente", cioè su porzioni ridotte delle stesse dette blocchi selezionati in modo aleatorio. Ciò equivale a generare un certo numero di ordinamenti casuali dove le impronte tendono a collocarsi ripetutamente entro una distanza contenuta rispetto ai propri simili. Tale proprietà consente di individuare i grappoli ovviando alla necessità di un confronto esaustivo spesso proibitivo in termini di tempo e di risorse. A seconda della tipologia delle chiavi identificative disponibili (nominativi e altri dati anagrafici, indirizzi e coordinate geografiche, schede tecniche di descrizione attività o prodotto, ecc..) è dunque possibile parametrizzare l'algoritmo (criteri di scomposizione delle stringhe in vettori/insiemi, lunghezza delle impronte, soglia di minima sovrapposizione affinché due record finiscano nello stesso grappolo, numero di ordinamenti casuali delle impronte) in modo tale da controllare l'errore atteso del primo (record erroneamente collocati all'interno di un certo grappolo) e/o del secondo tipo (record non riconosciuti come membri di un certo grappolo) a seconda delle esigenze o dei vincoli di qualità richiesti.

Obiettivi

L'obiettivo del progetto è sviluppare, testare e, infine, inserire nel toolkit RELAIS l'algoritmo MinHash (Broder, 1997) per la ricerca di record duplicati all'interno di banche dati integrate di grandi dimensioni in cui una parte non trascurabile di unità statistiche non è univocamente o sistematicamente identificabile. Data la complessità dell'operazione di integrazione tra fonti spesso concepite e sviluppate per fini diversi e contraddistinte da una qualità del dato non omogenea, poter aggiungere all'arsenale metodologico un algoritmo come MinHash consente una maggiore flessibilità ed un ventaglio più ampio di soluzioni alle sfide operative poste dal nuovo paradigma di produzione del dato statistico. Un altro ambito, oltre a quelli già citati, nei quali le metodologie sviluppate da questo Laboratorio potrebbero trovare terreno fertile di applicazione sono gli scanner data per le statistiche sui prezzi. Infatti la descrizione dei prodotti, sebbene

codificata, può presentare delle disomogeneità che rendono difficoltoso riconoscere casi in cui due o più record identificano di fatto lo stesso articolo.
A questo proposito è interessante confrontare il rendimento di MinHash, sia in termini di efficienza computazionale che di qualità dei grappoli ottenuti, non solo rispetto a SimHash ma anche rispetto alle tecniche più conosciute di record linkage probabilistico come, ad esempio, la procedura di Fellegi-Sunter.

Metodologia

Sono stati sviluppati in modo indipendente tre prototipi: uno principale in Java e altri due rispettivamente in Stata e in Python allo scopo di validare il primo. La scelta di utilizzare il codice sorgente Java è motivata dal successo con cui questo linguaggio era già stato impiegato nello sviluppo di SimHash in termini di prestazioni, affidabilità e compatibilità con RELAIS. Da segnalare che il confronto tra gli algoritmi non è puntuale perché la procedura utilizza permutazioni casuali che non è stato possibile replicare identiche nei tre linguaggi.
Il modulo principale Java usa l'algoritmo di permutazione casuale proposto da Daniel Dubnikov ed ha un parametro (seed) che garantire la ripetibilità degli esperimenti.
Gli algoritmi sono stati testati su unità statistiche di diversa natura come luoghi (indirizzi), individui (anagrafica) e scanner data (descrizione prodotto).

Risultati ottenuti

I risultati preliminari mostrano buone proprietà di scalabilità (i tempi di esecuzione aumentano linearmente sia rispetto al numero di record da confrontare sia rispetto alla lunghezza delle impronte). Il beneficio marginale in termini di nuovi duplicati identificati non sembra compensare il costo marginale (aumento dei tempi di esecuzione) oltre una certa lunghezza delle impronte e/o del numero dei blocchi di confronto. L'ottimizzazione di questi due parametri in base alla dimensione del problema di linkage o alla natura delle chiavi di confronto rimane oggetto di studio. MinHash è apparso non solo competitivo ma anche complementare rispetto a SimHash rivelandosi superiore in alcune applicazioni ma meno performante in altre.

Membri del team

Luca Mancini lmancini@istat.it
Luca Valentino luvalent@istat.it
Stefania Fatello fatello@istat.it
Stefano Daddi daddi@istat.it
Massimo De Cubellis decubell@istat.it
Alessandra Ronconi alronconi@istat.it