Universitá degli Studi di Milano Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

STAN: un tool per l'annotazione semantica incrementale di tabelle sul web

Relatore:
Dott. Matteo Palmonari

Co-relatore:
Dott. Riccardo Porrini

Relazione della prova finale di:
Davide Brando Preda
Matricola 735850

Anno Accademico 2014-2015
25/11/2015

Le tabelle sono un veicolo tramite il quale è possibile esprimere importanti informazioni che difficilmente sarebbe possibile esprimere in un testo libero. Le righe e le colonne di una tabella, infatti, conferiscono ai contenuti di essa una struttura chiara e precisa. Per un essere umano comprendere la semantica di una tabella può essere intuitivo anche proprio grazie alla sua struttura. Tuttavia lo stesso non si può dire di una macchina che si trova a dover interpretare una tabella presente sul web.
Le tabelle sono una delle poche pratiche di strutturazione dei dati sul web e probabilmente una delle più utilizzate. Infatti si conta che nei documenti web indicizzati da Google sono presenti oltre 154 milioni di tabelle contenenti dati relazionali. In questa tesi ci si concentra proprio sulla semantica dei dati contenuti nelle tabelle sul web, un problema che si inserisce in un contesto più generale che è l'interpretazione dei dati presenti nelle pagine web. La maggior parte delle informazioni contenute sul web infatti viene processata dai motori di ricerca sotto forma di semplici stringhe di testo non strutturate. In fase di ricerca dunque il discriminante per stabilire se un documento deve essere considerato rilevante o meno si basa fondamentalmente su metriche di frequenza e similarità dei termini cercati rispetto a quelli contenuti nei documenti. In questo modo però la semantica intrinseca nei documenti viene persa e questo impedisce ad una macchina di cogliere il significato dei contenuti di una pagina web.
Il Semantic Web affronta questa classe di problemi definendo degli standard per rendere i contenuti presenti sul web maggiormente strutturati. Il web infatti si è sempre basato sul linguaggio HTML che consente di strutturare una pagina web dal punto di vista grafico in modo che sia visualizzabile all'interno di un browser. L'HTML puro tuttavia non consente di strutturare i contenuti in modo che siano comprensibili da una macchina. Negli ultimi anni tuttavia si è diffuso l'utilizzo di annotazioni semantiche che, correttamente abbinate al linguaggio HTML, consentono di conferire una semantica non ambigua e condivisa a precisi elementi della pagina. Sfruttando dunque dei linguaggi di annotazione appositi come ad esempio RDFa è possibile arricchire i contenuti di una pagina con utili metadati. Queste informazioni aggiuntive sono poi estratte dai motori di ricerca e dai web crawlers per fornire agli utenti un'esperienza di ricerca più ricca ed esaustiva. Ad esempio Google sfrutta tali annotazioni per costruire automaticamente dei riquadri informativi a fianco dei consueti risultati di ricerca.

Annotare semanticamente un elemento significa arricchirlo in modo da assegnargli un significato non ambiguo e condiviso. In particolare l'annotazione mira ad assegnare a precisi elementi la semantica di un concetto o una proprietà contenuti in una ontologia. Un'ontologia è una rappresentazione formale, condivisa ed esplicita di una concettualizzazione di un dominio di interesse ed è interpretabile da una macchina. Dunque sfruttando le annotazioni e la conoscenza racchiusa nelle ontologie una macchina è in grado di interpretare i contenuti di una pagina web o i dati di una tabella. Una volta strutturati i contenuti diventa possibile integrarli

tra loro unendo pezzi di informazioni in relazione tra di loro. Conseguentemente all'integrazione dei dati è possibile fornire una risposta ad interrogazioni complesse che implicano la comprensione di dati provenienti da fonti eterogenee. Ad esempio, date due tabelle estratte dal web relative rispettivamente ai dati statistici sui terremoti avvenuti negli ultimi cento anni in Italia e i dati edilizi relativi al numero di edifici costruiti nelle città italiane, quello che si vorrebbe essere in grado di fare è trovare una correlazione tra questi eventi utilizzando le informazioni presenti in entrambe le tabelle. Per poter portare a termine questo compito è necessario comprendere la struttura e la semantica delle due tabelle, annotarle ed infine integrarle. Quello dell'annotazione di fatto è un problema noto nell'area di ricerca della Data Integration. Nella Data Integration classica, infatti, una delle tecniche utilizzate per l'integrazione prevede la creazione di un dataset integrato virtuale mediante la definizione di mapping tra gli schemi locali dei dataset da integrare e uno schema globale. Di fatto quindi l'annotazione può essere vista proprio come la definizione di un mapping verso uno schema globale che è rappresentato da una ontologia.

L'obiettivo di questa tesi dunque è quello di concentrarsi sui dati in formato tabellare ed implementare uno strumento che consenta di effettuare annotazioni semantiche su di essi. Il risultato di questa tesi, infatti, è STAN una applicazione web, accessibile all'indirizzo `http://stan.disco.unimib.it/`, che consente tramite interfaccia grafica di importare una tabella, annotarla con la semantica di una ontologia e infine esportare il risultato dei mapping definiti (figura 1).

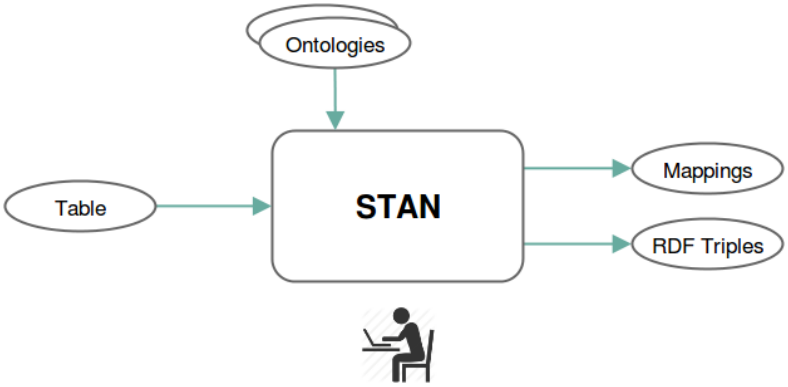

Figura 1: Il processo di annotazione di STAN.

L'implementazione del tool è stata guidata da due casi di studio reali attraverso i quali è stato possibile studiare e comprendere le problematiche relative all'integrazione dei dati. I due casi di studio fanno riferimento a due domini diversi e hanno fatto emergere esigenze e problematiche diverse. Il primo caso di studio è legato a temi socio-economici e tratta della decisione della città di Chicago di chiudere in massa 47 scuole della città per ridurre i costi di gestione. In questo ca-

so di studio esistono una molteplicità di sorgenti di dati in formato tabellare (e.g. crimini, rendimento delle scuole, etc) che dovrebbero essere integrate da esperti di dominio e poi analizzate per rispondere a domande complesse. Il secondo invece riguarda il mondo dell'eCommerce, e in particolare l'azienda 7Pixel, un'azienda italiana leader nel settore della comparazione di prezzi online e proprietaria dei siti di TrovaPrezzi.it[1] e Kirivo.it[2]. Nello specifico il caso di studio tratta il problema di gestire e integrare in modo dinamico offerte commerciali provenienti da fonti eterogenee che nella maggior parte dei casi arrivano all'azienda in formato tabellare. L'analisi di questi due casi di studio ha permesso di dedurre alcuni dei problemi e delle sfide principali che il tool implementato si pone di affrontare. In particolare i temi principali emersi sono la difficoltà nell'integrazione di dati eterogenei e nella loro interpretazione semantica, la necessità di automatizzazione del processo di integrazione nel caso di grandi volumi di dati e l'esigenza di rendere il processo eseguibile anche da persone che non possiedono competenze tecniche.

I contributi principali di questa tesi sono:

1. La formalizzazione del problema e in particolare la formalizzazione dei modelli possibili di annotazione di una tabella.
2. L'implementazione di un tool a supporto dell'annotazione di una tabella.
3. Una sperimentazione che mette a confronto l'algoritmo di annotazione usato nel tool e uno esistente da stato dell'arte.

A partire dall'analisi dello stato dell'arte e provando ad annotare le tabelle reali dei casi di studio è emersa l'esistenza di alcuni pattern ricorrenti nell'annotazione di un tabella. Una definizione rigorosa di questi pattern non è presente nello stato dell'arte, per questo uno degli obiettivi della tesi è la formalizzazione di una serie di modelli possibili di annotazione. I modelli quindi sono stati utilizzati con un duplice scopo: per definire in maniera non ambigua i tipi di grafo RDF che vengono prodotti mediante STAN e per condurre un confronto tra i diversi tool di annotazione disponibili.

STAN presenta alcune caratteristiche che lo differenziano dagli altri tool esistenti da stato dell'arte. Innanzitutto STAN, a differenza degli altri tool, consente di definire una propria ontologia in maniera incrementale durante il processo di annotazione e questo permette agli utenti di utilizzare una semantica propria tramite la definizione di nuove proprietà e delimitandone la semantica. Un'altra differenza tra STAN e gli altri tool è l'architettura puramente pensata per il web. STAN infatti è una applicazione web utilizzabile online che permette a ciascun utente di avere il suo spazio di lavoro personale. Gli altri tool invece sono pensati per essere scaricati ed utilizzati da un singolo utente sulla propria macchina. STAN inoltre

[1] `http://www.trovaprezzi.it`
[2] `http://www.kirivo.it`

è l'unico tool che sfrutta un algoritmo di annotazione automatica *instance based* basato sul confronto di un campione dei valori della colonna con i valori di migliaia di proprietà di una knowledge base. Infine STAN è l'unico tool che, sfruttando il proprio algoritmo di annotazione, espone un servizio di API pubbliche le quali possono essere interrogate da applicazioni di terze parti per effettuare l'annotazione automatica di gruppi di valori. Il servizio di API è stato pensato appositamente ragionando sul caso di studio di 7Pixel. Uno degli obiettivi della tesi, infatti, è quello di rendere STAN un provider di annotazioni all'interno dei sistemi di 7Pixel e le API sono il meccanismo più semplice per permettere l'integrazione con i sistemi legacy propri dell'azienda.

L'obiettivo della sperimentazione infine è quello di valutare l'algoritmo di annotazione in termini di qualità e di performance in relazione allo stato dell'arte. La sperimentazione è infatti stata condotta per studiare l'algoritmo sotto due diversi punti di vista: per valutarne la qualità dei suggerimenti nel caso specifico dell'e-Commerce e per valutarne le performance in termini di tempo. Sebbene l'algoritmo non sia un contributo della tesi, uno degli obiettivi è quello di integrarsi con i sistemi di 7Pixel e quindi la sperimentazione è necessaria per garantire all'azienda lo sviluppo di un tool di qualità. La sperimentazione dunque è stata condotta esclusivamente su un dataset composto da listini commerciali di proprietà dell'azienda utilizzando come verità le annotazioni effettuate in passato. Nel paragone con lo stato dell'arte si è scelto di confrontarsi con l'algoritmo utilizzato in Karma. Karma, infatti è uno dei pochi tool completi di Table Annotation con cui si confronta STAN, inoltre l'algoritmo implementato in Karma è l'unico il cui codice sia accessibile pubblicamente.
Dai risultati della sperimentazione è emerso che STAN e Karma si comportano in modo molto diverso a seconda che si trattino colonne di tipo testuale o numerico. Tra i due approcci Karma ottiene risultati migliori sulle colonne testuali, al contrario fa STAN per le colonne numeriche. Sebbene le colonne numeriche siano in numero inferiore rispetto a quelle testuali spesso nel dominio di 7Pixel si rivelano essere tra le più ambigue. Infatti, in molti casi risulta difficile distinguere colonne come prezzo, disponibilità e spese di spedizione quando il loro ordine di grandezza è molto simile. La sperimentazione è stata condotta senza adattare gli algoritmi al dominio specifico e questo dunque lascia ampi margini di miglioramento. In particolare gli scarsi risultati ottenuti da STAN nell'annotazione delle colonne di tipo testuale possono essere migliorati introducendo alcune euristiche specifiche di dominio. Ad esempio riconoscere alcuni pattern ricorrenti come "http://" aiuterebbe facilmente la corretta individuazione di colonne come *Link* e *Immagine* per le quali l'algoritmo attualmente ottiene risultati pessimi.

Lo sviluppo di STAN è stato pensato principalmente con due finalità: l'integrazione nei sistemi di 7Pixel e la diffusione come tool Open Source di annotazione. Al momento queste due anime dell'applicazione convivono e la logica che le differen-

zia viene specificata solamente in fase di deployment. Tuttavia, probabilmente in futuro queste due componenti potrebbero divergere sempre più con l'introduzione di nuove funzionalità e quindi gestire la logica che le differenzia solamente in fase di deployment potrebbe diventare troppo oneroso e complesso. Per questo in futuro si prevede di scindere STAN in due progetti esaltando individualmente le sue due anime. Nel caso di 7Pixel ci si concentrerà sul rendere l'algoritmo di annotazione più performante nel dominio dell'eCommerce introducendo alcune euristiche specifiche che consentano di ottenere suggerimenti di annotazione di maggior qualità. Al di fuori di 7Pixel invece si cercherà di promuovere STAN come strumento Open Source per l'annotazione di dati tabellari in formato Open Data e per questo ci si concentrerà sull'interfaccia grafica e sull'aggiunta di nuove funzionalità.