SAPIENZA
UNIVERSITÀ DI ROMA

Semantic Search App: strutturazione, correlazione e ricerca semantica dell'informazione multimediale

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

Corso di Laurea in Ingegneria Informatica e Automatica

Candidato

Matteo Mancanelli
Matricola 1711823

Relatore

Prof. Riccardo Rosati

Anno Accademico 2017/2018

Semantic Search App: strutturazione, correlazione e ricerca semantica dell'informazione multimediale

Tesi di Laurea. Sapienza – Università di Roma

Questa tesi è stata composta con LaTeX e la classe Sapthesis.

Email dell'autore: mancanelli.1711823@studenti.uniroma1.it

Dedicato a
Morten Tyldum

Indice

Capitolo 1

Introduzione

1.1 Panoramica

1.2 Dominio di interesse

Capitolo 2

Semantic Web

2.1 Informazione e Conoscenza

Il Web semantico viene definito per la prima volta in un articolo di Tim Berners-Lee del 2001, nel quale si parla di una estensione del World Wide Web in grado di potenziarne le capacità ed eliminare alcune delle intrinseche limitazioni. L'obiettivo è quello di spostare l'enfasi da una rete di documenti ad una rete di dati e rendere sempre più semplice esporre, connettere e condividere gli stessi tanto per gli utenti che ne fruiscono quanto per le applicazioni che li processano. Si intende perciò superare l'idea del Web come semplice archivio di documenti, caratterizzato da uno scambio poco flessibile, dall'information overload, da una assente cooperazione fra i diversi moduli software, da una mancata strutturazione gerarchica delle informazioni.

Molti dei problemi presentati sono dovuti al fatto che i dati sono pensati per essere utilizzati e fruiti direttamente dall'uomo, mentre un elaboratore non è in grado di conoscerne realmente il contenuto informativo. In questa ottica, il termine semantico assume sostanzialmente la valenza di elaborabile dalla macchina (machine understandable). Fornire significato ai dati vuol dire quindi associare delle informazioni utili perché una macchina possa manipolarli in base alla loro interpretazione.

Per quanto detto, una delle caratteristiche imprescindibili del Web semantico è l'accesso ad un insieme strutturato di informazioni e a un insieme di regole di inferenza da utilizzare per eseguire procedure di ragionamento automatico. È soprattutto quest'ultimo aspetto che rende la nuova ricerca sulla rete tanto vicina al più generale ambito della rappresentazione della conoscenza.

Il primo strumento utilizzato per realizzare il fine prefissato è quello di strutturare e corredare i documenti di metadati. Questi riflettono parte di quello che c'è da sapere su un certo insieme di dati e ciò che da tale conoscenza è possibile ricavare. Per esprimere i metadati si utilizzano i linguaggi di annotazione (o markup language) costruiti sulla base del tradizionale XML (eXtensible Markup Language). Le annotazioni sono quindi lo strumento elaborato per rappresentare il significato dei dati annotati.

Ai linguaggi di annotazione si aggiunge uno degli elementi chiave nel Web semantico, le ontologie. Una ontologia è "una rappresentazione formale ed esplicita

di una concettualizzazione di un dominio di interesse". In altre parole si parla di un sistema formale in cui è possibile esprimere enunciati e dedurre le loro conseguenze logiche in modo del tutto meccanico. Il ruolo delle ontologie emerge nel momento in cui i soli linguaggi di annotazione non sono in grado di legare i concetti e di stabilire le relazioni che intercorrono fra di essi. La più semplice relazione definibile fra due termini è la relazione di sussunzione (o is-a) nell'ambito della creazione di una tassonomia.

Il mezzo comunemente utilizzato per la definizione delle ontologie è rappresentato delle logiche descrittive. Queste sono utilizzate soprattutto per esprimere la conoscenza in termini di concetti e relazioni che li caratterizzano. I costrutti base utilizzati sono i predicati unari (concetti atomici), i predicati binari (ruoli atomici) e gli individui. Il fine è quello di fornire rigore formale e associare procedure automatiche di ragionamento. Una base di conoscenza comprende due componenti fondamentali, la prima per la definizione della conoscenza intensionale (TBox), utile nell'abito dei concetti e delle relazioni, e la seconda per la conoscenza estenzionale (ABox), utile nell'ambito degli oggetti presi in esame.

È evidente come l'uso di una ontologia rispecchia la necessità esplicitata di conoscere la semantica dell'informazione trattata ed eseguire inferenze per far emergere nuova conoscenza precedentemente celata. L'attività di inferenza permette così di estendere l'insieme degli assiomi asseriti in un determinato contesto, ma anche di verificarne la consistenza interna nel processo denominato validazione. Un ragionatore automatico (o reasoner) è un software in grado di svolgere tale attività su delle basi di conoscenza, anche per mezzo della logica dei predicati del primo ordine.

Per perseguire l'obiettivo di garantire ai dati un ruolo di primo piano nel nuovo paradigma, si introducono infine i concetti di vocabolari e Linked Data. Un vocabolario è una struttura condivisa per la rappresentazione delle informazioni che afferiscono ad uno stesso dominio. La possibilità di organizzare la conoscenza integrando i dati provenienti da diversi vocabolari ed eliminando le ambiguità con l'aiuto dei reasoner permette uno sviluppo indipendente e distribuito dei diversi domini, piuttosto che la creazione di un unica struttura poco modulare e complessa da processare.

Accanto all'uso di vocabolari e ontologie è necessario avere delle norme per rendere disponibili grandi quantità di dati e per far sì che possano essere trattati correttamente. Si parla quindi di Linked Data per riferirsi alle modalità di pubblicazione dei dati strutturati, utili per la successiva integrazione con altri dati preesistenti, l'inferenza e l'interrogazione semantica. I dataset più celebri sono quelli che compongono il progetto Linked Open Data (LOD), l'esempio meglio rappresentativo dell'utilizzo dei nuovi strumenti per la gestione dei dati annotati e per la definizione semantica degli stessi.

2.2 RDF, OWL e SPARQL

Numerose sono le tecnologie nate per il conseguimento dei diversi aspetti che coinvolgono la realizzazione del Web semantico. Fra queste si possono identificare le più popolari ed utilizzate per la definizione dei metadati, per la creazione delle ontologie e per le interrogazioni: RDF (Resource Description Framework), OWL (Web Ontology Language) e SPARQL (acronimo ricorsivo di SPARQL Protocol and RDF Query Language).

RDF è un semplice framework per la rappresentazione e lo scambio di metadati strutturati. Il modello concettuale che realizza RDF permette l'organizzazione delle informazioni in un insieme di triple, ognuna delle quali è composta da un soggeto, un predicato ed un oggetto. Il soggetto, ovvero la risorsa, e il predicato delle triple sono necessariamente rappresentati da un URI (Uniform Resource Identifier), mentre l'oggetto può essere un tipo di dato primitivo. Con RDF si definisce quindi un insieme di relazioni binarie che intercorrono tra due dati, ma non si forniscono regole di definizione per queste relazioni.

Un dataset RDF può essere visto e rappresentato come un digrafo, in cui i predicati sono gli archi orientati dai soggetti agli oggetti. Questa descrizione, adeguata nell'ambito della visualizzazione grafica, non è però adatta all'elaborazione automatica: si utilizza perciò una serializzazione testuale, espressa in diverse notazioni sintattiche fra cui RDF/XML, N-Triples e turtle (Terse RDF Triple Language). Alcuni elementi che possono essere utilizzati all'interno dello stesso RDF sono termini predefiniti, classi contenitore (o collezioni) e risorse anonime (blank nodes). Inoltre è possibile utilizzare una tripla RDF come soggetto di un'altra tripla, sfruttando il meccanismo della reificazione e creando così un nuovo livello di metadati.

RDFa (RDF in attributes) è una raccomandazione W3C per l'integrazione di annotazioni semantiche all'interno dei documenti HTML e XHTML. Si possono infatti aggiungere le triple RDF sfruttando un insieme di attributi definiti dal nuovo standard, in modo tale da arricchire il contenuto informativo delle pagine Web e renderle fruibili tanto agli utenti umani quanto agli agenti che processano i metadati. L'inserimento e la successiva estrazione di triple da un documento HTML si basa sul concetto di contesto: questo rappresenta il soggetto della tripla, al quale verranno associati tutti i valori dei successivi attributi fino al nuovo cambio di contesto.

Il primo passo verso la creazione di un linguaggio ontologico è quello di RDF-S (RDF Schema), pensato per aggiungere elementi che non possono essere espressi con il solo framework RDF. In particolare si possono definire relazioni tra termini generici invece che tra individui, con nuovi costrutti che classificano risorse e proprietà. Tali relazioni possono ancora essere rappresentate con un grafo orientato e viene mantenuta quindi la stessa sintassi utilizzata per RDF anche per gli altri linguaggi del Web semantico (approccio same-syntax). Una dichiarazione con RDF-S può essere vista come un vincolo sulla base di dati delle triple RDF. Una importante caratteristica delle dichiarazioni in RDF/RDF-S consiste nella possibilità di parlare di sé. Si parla in questo caso di proprietà di riflessività del linguaggio.

Una più completa ed espressiva famiglia di linguaggi ontologici è OWL: questo strumento è stato realizzato per estendere le possibilità concesse dagli standard

precedenti e costruire ontologie sempre più complesse, basate su definiti formalismi e sulle procedure di ragionamento automatico. La famiglia OWL si compone di tre differenti linguaggi: OWL-Lite, OWL-DL, OWL Full. Questi si differenziano per le possibilità che offrono nell'ambito della potenza espressiva. La ricerca di una maggiore espressività, e quindi di una più ampia capacità di costruzione delle ontologie, sacrifica però uno degli aspetti fondamentali del Web semantico, la decidibilità logica (o la trattabilità in termini di costo computazionale nel processo di reasoning).

A differenza di quanto accade in RDF-S, OWL permette di costruire una classe enumerandone gli individui o utilizzando gli operatori insiemistici come l'unione, l'intersezione o il complemento. Inoltre è possibile costruire proprietà con vincoli come quelli di quantificazione esistenziale, di quantificazione universale, di cardinalità minima e di cardinalità massima.

L'ultima delle tecnologie fondamentali nate per il supporto del Web semantico è SPARQL, un linguaggio di interrogazione utile per estrarre i dati codificati in basi di triple RDF. Questo strumento è stato reso standard nel 2008 dal Data Access Working Group, gruppo di lavoro del consorzio W3C. Così come SQL riflette, nella rappresentazione delle query, il modello relazionale su cui si applica, allo stesso modo SPARQL basa la propria rappresentazione sul concetto di grafo: una interrogazione viene espressa come il pattern di un grafo RDF e la risposta a tale interrogazione è costituita da tutte e sole le triple che istanziano quel pattern (graph matching).

Scendendo più in profondità nella struttura di una query SPARQL si possono idendificare diverse sezioni. Nella prima parte di dichiarano i prefissi utilizzati, si decide il tipo di risposta da ottenere (con i termini select, construct, ask e describe) e si definiscono i dataset su cui agire. A questo segue il vero e proprio graph pattern, costituito da un insieme di triple e di variabili per le quali si troveranno le corrispondenze nel dataset interrogato. Al termine si possono usare filtri e modificatori della query per ottenere un migliore controllo sulle informazioni presentate all'utente.

2.3 Modello Relazionale e Modello E-R

Il nuovo Web dei dati deve inserirsi all'interno di una ampia e complessa struttura di sistemi informativi largamente utilizzati nel mondo dell'elaborazione automatica. La manipolazione delle informazioni è comunemente affidata alla costruzione di sistemi di gestione delle basi di dati (DBMS). Nella strutturazione di una base dati si possono individuare tre fasi consecutive: la progettazione concettuale, la progettazione logica e la progettazione fisica. La prima fase, corrispondente alla progettazione concettuale, è essenzialmente realizzata sfruttando il modello E-R, strumento principale per la rappresentazione grafica in forma di schema o diagramma. Il modello relazionale è invece il modello più diffuso per la realizzazione della progettazione logica e per le interrogazioni SQL. È necessario quindi sviluppare tecniche per l'integrazione di questi modelli con le nuove tecnologie del Web semantico.

Il modello relazionale si basa sul concetto matematico di relazione. Detti $D_1, D_2, ..., D_n$ una serie di n domini, si definisce relazione un sottoinsieme di n-uple ordinate $(d_1, d_2, ..., d_n)$ generato dal prodotto cartesiano fra i domini $D_1 \times D_2 \times ... \times D_n$, dove $d_1 \in D_1, d_2 \in D_2, ..., d_n \in D_n$. Nel modello relazionale, la relazione è rappresentata in forma tabellare e ogni dominio è visto come un attributo a cui è associato un nome. La mancanza di una struttura posizionale rende l'ordinamento irrilevante, ma nello stesso tempo ogni tupla deve avere elementi di tipo omogeneo rispetto agli attributi definiti.

Considerato l'esteso e trasversale uso dei modelli relazionali sono diversi i tentativi di modellazione di metadati nelle basi di dati relazionali. Esempi di tali integrazioni sono le strutture miste, in cui si associa ad ogni attributo A un attributo A_c per inserire le annotazioni dei valori, e le associazioni intensionali, in cui le annotazioni vengono registrate in una relazione indipendente e poi associate alle tuple corrispiondenti per mezzo della valutazione di una query. Per lo storage e l'analisi di dati RDF è possibile creare un semplice sistema in cui, fra il DBMS e l'interfaccia utente, si inserisce un livello in grado di tradurre le interrogazioni RDF perché il sistema relazionale sottostante possa soddisfarle. Un metodo semplice ed efficace, anche se poco efficiente, per mappare un dataset RDF in una base relazionale è quello di creare una relazione con tre attributi (soggetto, predicato ed oggetto) e costruire le tuple con le triple del dataset.

Nel modello E-R la struttura dello schema concettuale è descritta in forma grafica e definisce il livello intensionale della base di dati. I costrutti fondamentali utilizzati per tracciare i diagrammi E-R sono le entità, le relazioni, i vincoli e gli attributi. Le entità rappresentano le classi degli oggetti di interesse mentre le relazioni sono associazioni che legano due o più entità. Una base di dati, così come una ontologia, è un metodo per rappresentare un insieme di informazioni: sembra quindi immediata, almeno dal punto di vista speculativo, la traduzione della struttura definita da uno schema E-R in un dominio ontologico che sia fedele alle regole imposte dal Web semantico. Questa operazione è in effetti possibile, sebbene esistano delle limitazioni e sia necessario tenere in considerazione le numerose differenze che intercorrono fra un diagramma e un linguaggio come OWL.

La prima divergenza che può essere notata è l'ipotesi di mondo chiuso, assunzione secondo la quale una dichiarazione è vera se e solo se è effettivamente definita ed è

considerata falsa se il suo valore di verità non è noto. Questa ipotesi caratterizza generalmente i modelli relazionali. Al contrario, OWL opera l'assunzione opposta, l'ipotesi di mondo aperto, in cui è ammessa una conoscenza incompleta dei dati di interesse. La stessa divergenza si trova sull'ipotesi di unicità del nome: l'idea per cui due nomi diversi si riferiscano ad entità diverse non è accettata a priori in OWL.

Dal punto di vista dei costrutti, le entità di un diagramma E-R possono essere direttamente tradotte nelle classi OWL, mantenendo la struttura tassonomica con l'uso di superclassi e sottoclassi. Diversi generi di vincoli devono essere aggiunti per determinare le restrizioni sugli individui che possono appartenere ad una classe. È importante notare che, alla pari dello schema E-R, le sottoclassi ereditano le proprietà delle rispettive superclassi. Le relazioni binarie fra entità del diagramma sono tradotte nelle Object Properties, proprietà che legano individui di due classi (non necessariamente distinte), mentre gli attributi delle entità si traducono in Datatype Properties, che legano un individuo a un tipo di dato primitivo (integer, float, string...). In OWL è possibile definire il dominio ed il range delle proprietà ed aggiungere un set di restrizioni di quantità e cardinalità per gli individui che sono coinvolti dalle proprietà stesse. Non è possibile al contrario modellizzare fedelmente gli attributi delle relazioni e soprattutto definire le chiavi primarie di una classe, presenti e fondamentali nel modello relazionale.

Capitolo 3

Vector Space Model

3.1 Sistemi di Raccomandazione

I sistemi di raccomandazione sono software di manipolazione del contenuto informativo, largamente utilizzati per gestire grandi quantità di dati e selezionare gli stessi in funzione di una migliore fruizione da parte degli utenti a cui sono rivolti. Questi sistemi vengono impiegati in modo evidente soprattutto dalle piattaforme online che operano in ambito e-commerce, come Amazon, e streaming on demand, fra cui Netflix e Spotify. Un sistema di raccomandazione si rende perciò indispensabile in tutte le occasioni in cui è necessario contenere il volume delle informazioni e creare un metodo di filtraggio per la visualizzazione e la presentazione. L'obiettivo: aiutare le persone ad effettuare scelte in base alle proprie preferenze. In modo più formale: detto C l'insieme degli utenti e S l'insieme degli elementi, il migliore elemento per un utente è quello che massimizza la funzione di utilità $u(c, s)$.

$$\forall\, c \in C, \quad \hat{s}_c = \underset{s \in S}{\operatorname{argmax}}\ u(c, s)$$

La prima fase di cui si compone un sistema di raccomandazione è la raccolta dei dati relativi agli oggetti di interesse per l'applicazione e alla profilazione degli utenti che ne usufruiscono. Il sistema migliorerà il servizio offerto solo nel momento dispone di sufficiente conoscenza: questa fase è quindi di fondamentale importanza per il conseguimento di buoni risultati, indipendentemente dalla tecnica di filtraggio messa in atto. Per quanto detto emerge in modo chiaro la stretta relazione che intercorre fra i sistemi di raccomandazione e l'attuale ambito definito Big Data Analysis. Gli input su cui lavorerà il sistema possono essere forniti esplicitamente (ad esempio, attraverso un metodo di rating) o, in modo più sofisticato, ottenuti mediante l'osservazione del comportamento degli utenti ed una successiva deduzione delle loro preferenze. Quest'ultimo metodo resta, allo stato dell'arte, meno performante del precedente, anche se rappresenta al meglio l'idea sottostante un sistema di raccomandazione ed è immune dal problema del bias.

In base alla tecnica usata per ottenere le raccomandazioni, i sistemi possono essere classificati in quattro categorie: filtraggio content-based, filtraggio collaborativo, filtraggio ibrido e filtraggio semantic-social. Le tecniche content-based sfruttano gli attributi degli oggetti di interesse per ottenere i risultati piuttosto che le informazioni

sugli utenti. Questo criterio è principalmente utilizzato nell'analisi di documenti testuali e pagine web ed estende alcuni degli approcci tipici dell'information retrieval. Si impiegano a tal proposito metodi quali Vector Space Model e Neural Networks per estrarre relazioni fra i documenti e fornire le raccomandazioni agli utenti. Il principale svantaggio risiede nella necessità di una conoscenza approfondita delle feature (caratteristiche) sulle quali fondare la valutazione. Nello stesso tempo però non si richiede la profilazione degli utenti e la condivisione delle loro informazioni, favorendo così la protezione della privacy.

Il metodo collaborativo crea dei suggerimenti utilizzando la similarità tra gli utenti. Si registrano le scelte del maggior numero di persone e si forniscono raccomandazioni associandole in gruppi definiti neighborhood. Utenti con stesse preferenze possono quindi scambiare indirettamente diverse informazioni fra loro, suggerendo nuovi contenuti attraverso il sistema. In questo caso, l'utilità $u(c, s)$ dell'elemento s per la persona c è basata sull'utilità $u(c', s)$ assegnata ad s dalle persone $c' \in C' \subseteq C$ che sono simili a c. Anche questo metodo, nonostante sia più dinamico e attuale, non è esente da particolari svantaggi, primi fra tutti la scalabilità, la sparsità dei dati e la mancanza di sufficienti informazioni sugli utenti che si autenticano per la prima volta o che non condividono il loro profilo.

Le tecniche usate per la realizzazione dell'approccio collaborativo sono svariate e spesso mutuate dai settori del machine learning e del data mining (model-based collaborative filtering). Si parla quindi di modelli probabilistici e algoritmi di apprendimento come le reti neurali, ma anche regole di associazione, alberi di decisione e clustering. In questi algoritmi, come per la funzione utilità, il rating $r_{c,s}$ della persona c per l'oggetto s viene calcolato sulla base dei rating $r_{c',s}$ delle persone più simili a c. In particolare è possibile usare una media pesata dei rating rispetto a tutti gli utenti c' del tipo

$$r_{c,s} = \widetilde{r}_c + k \sum_{c' \in C'} sim(c, c') \cdot (r_{c',s} - \widetilde{r}_{c'})$$

dove

- k è un fattore normalizzante del tipo $k = 1/\sum |sim(c, c')|$
- $\widetilde{r}_c$ è il rating medio di una persona, $\widetilde{r}_c = (1/|S_c|) \sum_{s \in S_c} r_{c,s}$
- S_c è l'insieme degli elementi relativi a c, ovvero $S_c = \{s \in S \,|\, r_{c,s} \neq 0\}$
- $sim(c, c')$ è la similarità fra c e c'

Nel tentativo di superare le criticità ad ottenere il miglior sistema di raccomandazione è stato sviluppato l'approccio ibrido, che integra i metodi descritti in precedenza e li sintetizza con strategie differenti. Infine, negli ultimi anni, sono nati nuovi sistemi ispirati al crescente utilizzo dei social network. L'approccio semantic-social si fonda perciò sulla modellazione delle reti sociali e delle interazioni fra persone ed oggetti. Le reti sono rappresentate da strutture astratte come i grafi e possono essere visitate con i tradizionali algoritmi di ricerca su grafo come depth-first search (DFS).

3.2 TFIDF e Cosine Similarity

Capitolo 4

Data Analysis e UI

Capitolo 5

Metodologia

Capitolo 6

Implementazione

6.1 Kaggle e ER

6.2 Protege e Python

6.3 Java con query e correlazione

6.4 Node.js con JQuery per paginazione e ricerca

6.5 Prestazioni

Capitolo 7

Conclusioni

7.1 Risultati Ottenuti

Argomenti

1. Obiettivi e strumenti
2. Dominio di interesse
3. Dati Utilizzati (Kaggle) e manipolazione preliminare
4. Diagrammi ER e ontologico
5. Creazione Ontologia OWL (Protege)
6. Description Logics
7. Triple RDF da csv (Python e rdflib)
8. Vector Space Model
9. Correlazione (Java e Jena)
10. Query semantiche (da SQL a SPARQL)
11. Rappresentazione Web (UI)
12. Node.js e Express.js
13. Paginazione e ricerca (JQuery)
14. Data Analysis
15. Confronto prestazioni
16. Reverse Engineering e ottimizzazione delle prestazioni
17. Risultati ottenuti