Linked Open Data for content-based recommender systems

Luciano Quercia
Simone Rutigliano

April 18, 2013

1 Introduzione

Il nostro progetto consiste nella realizzazione di un content-based recommender system che raccomandi film utilizzando dati provenienti dalla Linked Open Data Cloud al fine di poter aumentare l'efficienza del recommender system utilizzando informazioni aggiuntive inerenti un particolare film utilizzando differenti fonti.

1.1 Linked Data

L'interoperabilità è uno dei vantaggi più importanti del modello Open Data. I dati, se isolati, hanno poco valore; viceversa, il loro valore aumenta sensibilmente quando data set differenti, prodotti e pubblicati in modo indipendente da diversi soggetti, possono essere incrociati liberamente da terze parti. Questo è alla base del processo di creazione di valore aggiunto sui dati: le applicazioni. Le applicazioni, di valore sociale e/o economico, sfruttano quello che può essere visto come un grande database aperto e distribuito per offrire viste e servizi. L'interoperabilità è dunque un elemento chiave di uno degli aspetti più innovativi offerti dagli open data: l'uso dei dati in modi e per scopi "inattesi", nuovi in quanto non previsti dai singoli enti e soggetti che pubblicano i "dati grezzi".

Per consentire il riuso dei dati occorre poter combinare e mescolare liberamente i dataset. Occorre cioè collegare i dati tra loro, stabilendo un link diretto quando i dati (possibilmente provenienti da diverse sorgenti) si riferiscono a oggetti identici o comunque relazionati tra loro. Tale collegamento diretto si manifesta come la possibilità di "saltare" da un dataset all'altro, ad esempio quando si vuole accedere a dati (come i dettagli su una particolare entità) che non si posseggono al'interno. Supponiamo per esempio di avere, da una parte, amministrazioni locali che pubblicano dati aperti relativi ai monumenti storici e agli hotel che si trovano nelle vicinanze di quei monumenti; dall'altra, Sovrintendenze ai Beni Culturali che pubblicano dati dettagliati sui monumenti, gli artisti e i periodi storici, e sui quadri esposti nei musei o nei palazzi. Combinare i due dataset potrebbe essere di grande utilità, ad esempio per offrire un servizio personalizzato sugli itinerari in base agli interessi culturali specifici di un turista.

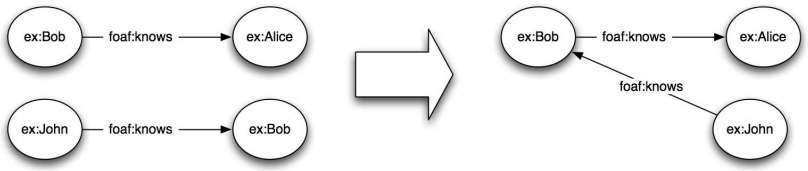
Per fare questo, se i dati non sono "collegati" (linked) occorre in qualche modo creare questi link, processando i dati a mano o attraverso algoritmi ad hoc. Questo processo può non essere banale e sicuramente è una barriera al riuso organico dei dati.

Nei cosiddetti Linked Data, questi collegamenti e relazioni tra le entità descritte nei dataset sono espliciti.

1.2 Machine readable vs. machine linkable

I linked data, per definizione, vengono espressi tramite Resource Description Framework (RDF). RDF non è propriamente un formato di dati, ma un "data model", cioè un formalismo per rappresentare dati. Un dataset RDF può essere infatti serializzato in diversi formati (RDF/XML, N3, NTriple, etc.), ma il data model RDF possiede alcune caratteristiche che restano immutate, a prescindere dal formato che viene utilizzato.

In poche parole il modello RDF è costituito da triple, della forma soggetto-predicato-oggetto. Le triple possono condividere oggetto o soggetto così da formare un grafo.



Questo insieme di triple RDF (o grafo) può essere espresso, allo scopo di essere scambiato tra applicazioni e pubblicato sul web, in vari formati di serializzazione.

Ad esempio in RDF/XML:

```
1  <?xml version="1.0"?>
2  <rdf:RDF
3     xmlns:ex="http://example.org/"
4     xmlns:foaf="http://xmlns.com/foaf/0.1//"
5     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
6     <rdf:Description rdf:about="http://example.org/John">
7        <foaf:knows>
8           <rdf:Description rdf:about="http://example.org/Bob">
9              <foaf:knows rdf:resource="http://example.org/Alice" />
10          </rdf:Description>
11       </foaf:knows>
12    </rdf:Description>
13 </rdf:RDF>
```

La caratteristica più importante di tale modello, che si sposa con la visione Linked Data, è usare Uniform Resource Identifier (URI).

2 Related Works

Mirizzi et al. (1), affrontando lo stesso problema, hanno utilizzato un VSM puro trasformando il grafo RDF di partenza in un tensore 3-dimensionale di adiacenza.

Passant (2), invece, in un contesto differente dal nostro (artisti musicali per poterle confrontare con le nostre

3 Misure

Il nostro recommender system è stato testato su

Passant D Misura di passant diretta; Indica il numero di archi diretti che collegano i film a e b Considerato che il grafo preso in considerazione è orientato, vanno presi sia gli archi che vanno da a a b, sia gli archi che vanno da b ad a. Cd(n,ra,rb) restituisce il numero di archi che vanno dal film a al film b. Cd(n,rb,ra) restituisce il numero di archi che vanno dal film b al film a.

Passant I Indica la distanza indiretta trai film presi in esame; il valore restituito sta ad indicare il numero di archi che mettono in relazione i due film passanti attraverso delle risorse condivise.

Cio(n,ra,rb) indica il numero di archi che soddisfano la seguente proprietà: data una tripla RDF: soggetto – predicato – oggetto, si sommano gli archi contenuti nelle triple rdf aventi come soggetti i due film in questione e come oggetto un film in comune ad entrambi le triple rdf. ra – predicato – oggetto1 rb – predicato – oggetto1 Cii(n,ra,rb) indica il numero di archi che soddisfano la seguente proprietà: data una tripla RDF: soggetto – predicato – oggetto, si sommano gli archi contenuti nelle triple rdf aventi come oggetti i due film in questione e come soggetto un film in comune ad entrambi le triple RDF. oggetto1 – predicato – ra oggetto1 – predicato – rb

Passant C Indica la distanza combinata tra i due film in questione utilizzando sia la distanza indiretta, che quella diretta.

Nostra Si prendono in considerazione sia gli archi diretti che uniscono a e b, si gli archi indiretti, cioè quelli che legano a e b ad un oggetto o soggetto comune. Un link è un collegamento diretto tra due film con un suo peso dato dalla produttoria dei pesi degli archi precedentemente trovati. La distanza coincide con la sommatoria del peso di ogni link.

PROFILI Simple Il profilo coincide con l'insieme dei film valutati positivamente (4 e 5 considerati in egual misura) Simple negative I film valutati 4 e 5 sono nell'insieme dei positivi. I film valutati 1 e 2 sono nell'insieme dei negativi. Voted Nostra Il profilo è l'insieme dei film visti, pesati secondo la formula: votazioni dei film visti - voto medio dei film. Voted Musto Il profilo è l'insieme dei film visti, pesati secondo la formula: Voto massimo + 1 - voto del film visto.

$$LDSD_d(r_a, r_b) \quad = \quad \frac{1}{1 + C_d(n, r_a, r_b) + C_d(n, r_b, r_a)} \tag{1}$$

Figure 1: LDSD Distance

4 Progetto

5 Sperimentazione

References

[1] R. Mirizzi, T. Di Noia, V. C. Ostuni, and A. Ragone. Linked open data for content-based recommender systems. 2012.

[2] A. Passant. Measuring semantic distance on linking data and using it for resources recommendations. In *Proceedings of the AAAI Spring Symposium" Linked Data Meets Artificial Intelligence*, volume 3, 2010.