

# 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

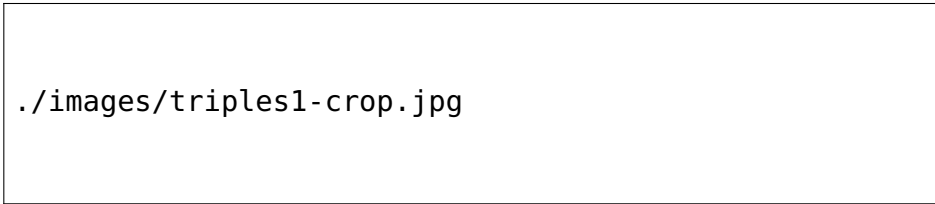
particolare entità) che non si posseggono all'interno. Supponiamo per esempio di avere, da una parte, amministrazioni locali che pubblicano dati aperti relativi ai monumenti storici e agli hotel che si trovano nelle vicinanze di quei monumenti; dall'altra, Sovrintendenze ai Beni Culturali che pubblicano 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, NTriples, 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.



```
./images/triples1-crop.jpg
```

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:

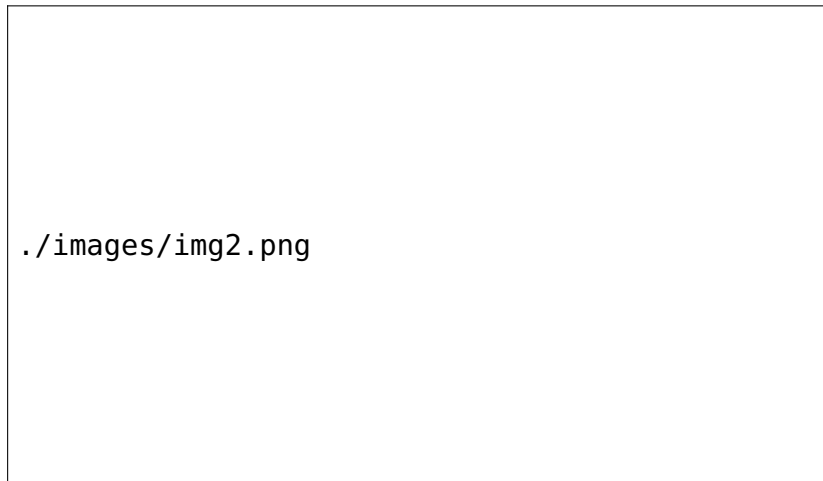
---

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

Tornando all'esempio del monumento, supponiamo che i due dataset (amministrazione locale e sovrintendenza) siano stati pubblicati come Linked

Data. Per identificare i monumenti, il dataset delle sovrintendenze usa URL (del tipo *http://cultural-heritage-example.org/monument/XYZ*). Il contenuto digitale di tali URL corrisponde alla descrizione dettagliata dei monumenti. Il data set dell'amministrazione locale, inserendo dei link a tali URL, come avviene in figura 1, permetterebbe a un software di risolvere l'URL e ottenere la descrizione del monumento (sempre aggiornata).

Ancora, dal momento che RDF consente di specificare precisi tipi di risorse, potremmo pensare a un semplice script che trovi tutte le risorse di tipo "monumento" nel dataset dell'amministrazione locale, e che importi, per ciascuna, informazioni aggiuntive, creando così un dataset misto. Su quest'ultimo nuovodato set arricchito, si potrebbero poi fare query del tipo "trova tutti gli alberghi vicini a un monumento successivo al XIII secolo, in cui siano esposte sculture del Canova".



Questo esempio è solo uno degli scenari possibili in cui i linked data possono favorire l'interoperabilità tra dataset. Le possibilità sono infinite se pensiamo alla vasta quantità di Linked Open Data già presenti sul Web. **DB-Pedia.org**, per esempio, espone una grande porzione di dati di Wikipedia come linked data, mentre **Geonames** offre descrizioni RDF di entità geografiche. *http://linkeddata.org/* fornisce un quadro dello stato corrente della "Linked data cloud", e mostra un ecosistema di database interconnessi in rapida crescita.

## 2 Lavori correlati

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 Progetto

### 4 Sperimentazione

### 5 Misure

#### 5.1 Distanze

Il nostro recommender system è stato testato su quattro diverse misure di distanza. Le prime tre sono descritte in Passant (2), la quarta è stata realizzata *ex novo*.

##### 5.1.1 Passant Direct

La distanza diretta. Indica il numero di archi diretti che collegano i film  $a$  e  $b$ . Tenendo conto che il grafo preso in considerazione è orientato, vanno presi gli archi in entrambe le direzioni.

In maniera formale,  $C_d(n, r_a, r_b)$  restituisce il numero di archi che vanno dal film  $a$  al film  $b$ .  $C_d(n, r_b, r_a)$  restituisce il numero di archi che vanno dal film  $b$  al film  $a$ .

La formula della distanza completa sarà quindi:

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

##### 5.1.2 Passant Indirect

Indica la distanza indiretta tra i 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.

Formalmente, dati:

$$C_{io}(n, r_a, r_b) = |\{o \mid (r_a \text{ } p \text{ } o) \wedge (r_b \text{ } p \text{ } o)\}|, \text{ con } p \in P$$

$$C_{ii}(n, r_a, r_b) = |\{s \mid (s \text{ } p \text{ } r_a) \wedge (s \text{ } p \text{ } r_b)\}|, \text{ con } p \in P$$

la formula indiretta di Passant è:

$$P_i(r_a, r_b) = \frac{1}{1 + C_{io}(n, r_a, r_b) + C_{ii}(n, r_b, r_a)} \quad (2)$$

indica il numero di archi che collegano indirettamente una coppia di triple 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 entrambe le triple RDF.  $r_a$  - predicato -  $r_b$  - predicato -  $C_{ii}(n, r_a, r_b)$  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 -  $r_a$  oggetto1 - predicato -  $r_b$

### 5.1.3 Passant Combined

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.

## 5.2 Profili

### 5.2.1 Simple

Il profilo coincide con l'insieme dei film valutati positivamente (4 e 5 considerati in egual misura)

### 5.2.2 Simple Negative

I film valutati 4 e 5 sono nell'insieme dei positivi. I film valutati 1 e 2 sono nell'insieme dei negativi.

### 5.2.3 Voted Nostra

Il profilo è l'insieme dei film visti, pesati secondo la formula: votazioni dei film visti - voto medio dei film.

#### 5.2.4 Voted Musto

Il profilo è l'insieme dei film visti, pesati secondo la formula:  $\text{Voto massimo} + 1 - \text{voto del film visto}$ .

## 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.