UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

Relazione

Un osservatorio sul turismo:
Il caso di Tourpedia

RELATORI:

Dott. Andrea Marchetti

Ing. Angelica Lo Duca

CANDIDATO:

Pietro Gianluca Calamia

Anno Accademico 2014-2015

Indice

Abstract

La rete e l'innovazione tecnologica sono elementi di estrema importanza in un momento di forte cambiamento culturale e sociale, che coinvolge direttamente tutta la Travel Industry.

In Italia ci sono 20 milioni di utilizzatori di smartphone e il 31% di coloro che ne possiedono uno ha prenotato almeno un hotel tramite mobile; il 29% ha cercato un ristorante, il 26% ha richiesto e ottenuto informazioni sul proprio viaggio, il 20% legge frequentemente recensioni su strutture turistiche.

L'evoluzione tecnologica mette tutti di fronte a nuove sfide e il settore dei viaggi e del turismo deve cogliere i segnali di questa trasformazione in atto e rispondere adeguatamente alla domanda del mercato. Oggi chi viaggia per affari o per piacere vuole avere a disposizione il maggior numero di informazioni possibili in tempo reale e dà molta rilevanza all'esperienza condivisa in rete da altri. Il 78% dei viaggiatori, infatti, raccoglie sul web le informazioni necessarie per la propria vacanza e il 72% degli utenti di social network le utilizza mentre è in viaggio

In questo lavoro di tesi è descritta una specifica procedura utilizzata per estendere Tourpedia, un dataset geografico contenente dati turistici. Il dataset è formato da circa cinquecentomila luoghi estratti da quattro social media: Facebook, Foursquare, GooglePlaces and Booking.com. In dettaglio, sono stati aggiunti a tutti i luoghi, nello specifico agli alberghi, tutte le attrazioni e i punti di interesse nelle vicinanze.

Tourpedia è dotata anche di un'applicazione web, la quale permette di vedere tutti i luoghi sulla mappa, capire se un albergo è buono oppure no tramite un'analisi sentimentale, trovare e calcolare il percorso dall'albergo a un'attrazione vicina e vedere che servizi offre l'albergo.

Per arricchire Tourpedia sono state usate diverse tecniche di NLP (Natural Language Processing) usando una procedura di disambiguazione e di filtraggio in modo da sistemare meglio i nuovi dati ottenuti. Inoltre è stata cambiata l'interfaccia grafica all'applicazione web in modo da renderla più moderna ma anche più intuitiva.

Tramite Tourpedia è possibile vedere quali sono le attrazioni o i punti di interesse vicini ad un albergo, infatti tramite tecniche di analisi linguistica è possibile stabilire una relazione tra due luoghi. Le relazioni tra alberghi e attrazioni turistiche potrebbero promuovere diverse ricerche nel settore dei servizi. In effetti esistono molti studi in letteratura che mostrano come gli alberghi scelgono la loro posizione in una città e come i turisti cambiano in base alla posizione di una struttura.

La mia tesi fornisce una panoramica di come Tourpedia sfrutta le basi di dati riguardanti una serie di alberghi e strutture e descrive come è possibile connetterle in modo da ottenere nuove informazioni utili a capire quali sono le abitudini dei turisti.

1 Introduzione

L'iniziativa dei Linked Data ha portato nuove opportunità per la costruzione di applicazioni Mashup Semantiche che chiameremo Linked Data Mashup (LDM). Dall'esposizione di dataset precedentemente isolati come ad esempio dei grafici di dati, che possono essere interconnessi e integrati con altri insiemi di dati, i Linked Data permettono la creazione di una scala globale interconnessa di spazio di dati, noto principalmente come il "Web dei dati". D'altra parte, i social media e le tecnologie mobili stanno permettendo quello che viene chiamato "database crowdsourcing" che permette le estrazioni di luoghi, eventi, recensioni ecc. che hanno un enorme potenziale di arricchimento nelle applicazioni mashup.

Un mashup Linked Data è un'applicazione (web) che permette nuove funzionalità estraendo e trasformando i dati già presenti sul Web. Potenziate da strumenti e tecnologie che sono state sviluppate dalla comunità semantica del Web le applicazioni Linked Data Mashup hanno permesso di analizzare in un nuovo modo i dati provenienti dal web.

1.1 Il problema del turismo

Il turismo è un fenomeno in continua espansione ed evoluzione. Oggi grazie ai nuovi mezzi di trasporto, è possibile visitare qualsiasi parte del globo con semplicità ed efficienza.

Il turismo di massa nacque negli Stati Uniti negli anni '50 e fu favorito dalla crescita del trasporto aereo intercontinentale. In poche ore di volo si potevano raggiungere luoghi e mete lontane. Il boom economico statunitense si replicò dieci anni dopo in Europa. I mercati di massa e la sindacalizzazione del lavoro favorirono la crescita del potere di acquisto minimo e rendendo il prezzo dei viaggi sempre più abbordabili per tutti. L' aerea industriale cominciò a differenziare l'offerta introducendo i primi voli charter mentre il progresso rese più rapida e immediata la trasmissione radio-tv da ogni luogo del mondo. Oggi la televisione e Internet svolgono un ruolo fondamentale nella comunicazione del turismo.

Grazie a sofisticate tecnologie, rispetto al passato, oggi è molto più veloce ed agevole l'organizzazione di un viaggio poiché esistono in rete portali predisposti per

l'organizzazione di viaggi e la ricerca di servizi turistici. Negli ultimi anni ha avuto una forte crescita la vendita dei cosiddetti viaggi last minute, pacchetti di viaggio acquistati pochi giorni prima della partenza. Inoltre, lo sviluppo delle compagnie aeree low cost, nate all'inizio degli anni novanta, hanno dato forte impulso ai viaggi di breve durata in ogni periodo dell'anno.

Secondo uno studio cinese, fatto da Yang e Kevin K.F. Wong, la scelta di un albergo dipende da più fattori come l'accessibilità, effetto agglomerazione, beni e servizi pubblici, sviluppo urbano, costo, classificazione, proprietà e servizi offerti.

Nell'ambito dell'informatica umanistica il problema del turismo è fondamentale, infatti il problema del turismo si potrebbe benissimo collocare a metà tra le scienze informatiche e le discipline umanistiche. L'informatica è necessaria per fornire al turista gli strumenti necessari per la navigazione web, mentre le discipline umanistiche studiano come è possibile migliorare i servizi necessari ad un turista affinché si possa migliorare la qualità del servizio.

Il congiungimento delle due aree dovrebbe mettere a disposizione al turista uno strumento completo in grado di eseguire i compiti richiesti.

1.2 Dal turismo di élite al turismo di massa

La complessità del turismo, sia come oggetto di studio nelle teorizzazioni scientifiche, sia come campo di applicazione nella pratica, è dovuta al suo essere un settore trasversale e, quindi, fattore di spinta e d'impatto di una molteplicità di effetti prodotti da settori differenti. Da tale complessità, discende una letteratura, improntata su metodologie e logiche diverse, in cui sociologi, antropologi, geografi, letterati, economisti, hanno descritto il ruolo del turismo nell'evoluzione della società contemporanea, ponendo anche le basi per la conoscenza del relativo settore economico.

Secondo l'UNWTO (*United Nations Word Tourism Organization*), per turismo si intende «*il movimento di persone che si spostano dal luogo di residenza ad un altro luogo, dove si fermano per tempo libero o per affari per almeno una notte*».

Si tratta di una definizione che, sebbene attiene al turismo moderno, ben contempla gli elementi propri del fenomeno turistico fin dalle sue origini (lo *spostamento*, la *durata*, la *motivazione*…), e che lascia intuire la difficoltà, non solo di definire che cosa sia il «turismo» (molteplici sono le definizioni che i diversi autori ne hanno dato negli anni), bensì di quantificare tale fenomeno: la mobilità del collettivo statistico (turisti) sul territorio, il regime di libera circolazione delle persone alle frontiere, nonché le nuove "modalità di fare turismo" (il turismo residenziale, il caravanning, ecc.), rendono sempre più complessa la rilevazione statistica dei flussi turistici.

Nonostante queste difficoltà, i dati raccolti confermano il costante trend di crescita che ha caratterizzato il turismo a partire, in particolare, dagli anni Novanta con un aumento medio degli arrivi internazionali di circa il 5% ed il coinvolgimento di nuove aree.

Secondo stime internazionali, all'indomani della Seconda Guerra Mondiale, il movimento interessava all'incirca 25 milioni di persone, divenute 70 milioni già nel 1960. Nel 1995, il movimento turistico internazionale ha superato i 500 milioni di arrivi per un fatturato di oltre 300 miliardi di dollari. Interrotto temporaneamente in seguito alla crisi provocata dai tragici fatti dell'11 settembre 2001, esso è ripreso gradualmente nel corso del 2002, fino a determinare, alla fine dello stesso anno, secondo i dati dell'OMT (Organizzazione Mondiale del Turismo), un nuovo aumento del 3,3% dei flussi turistici, che ha permesso di superare per la prima volta la quota di 700 milioni di arrivi.

Una vera e propria esplosione si è avuta nel 2004 con un aumento degli arrivi a livello internazionale di oltre il 10%. Le stime dell'OMT prevedono che saliranno ad oltre un miliardo, per superare il miliardo e mezzo nel 2020, con una ridistribuzione dei flussi che tende a privilegiare l'area Pacifico-Asia Orientale (dal 14,4% attuale al 25,4%) a discapito dell'Europa, la cui quota di mercato si prevede ridursi dal 59,8% al 45,9% (vedi fig. 1).

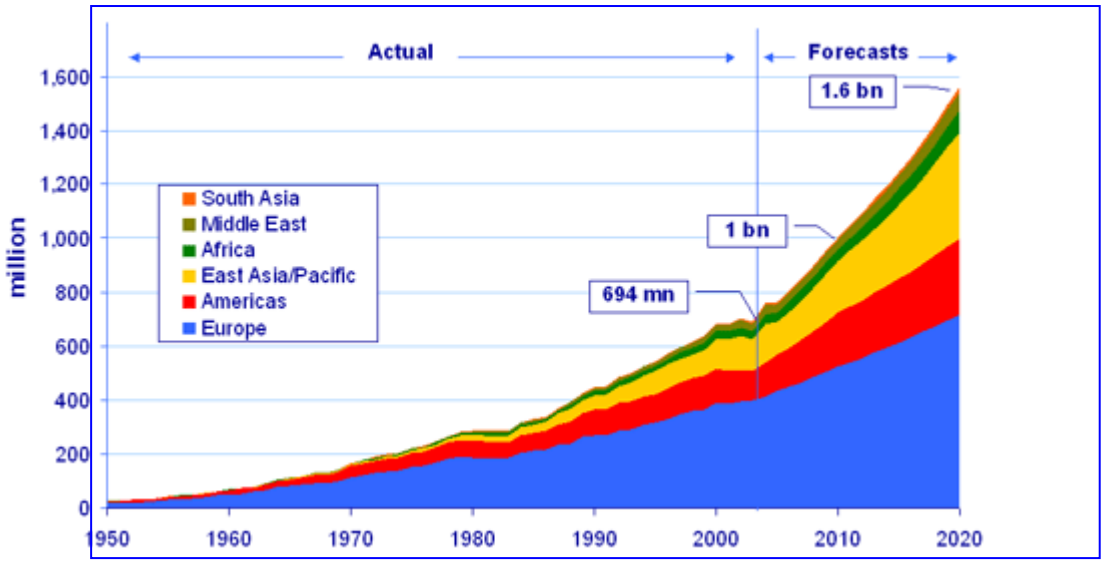

Fig. 1 - Previsioni di lungo termine dell'andamento dei flussi turistici internazionali

1.3 Quali sfide lancia il turismo oggi

Ai giorni nostri la tecnologia e i social media hanno completamente cambiato modo di vedere il turismo. Oggi abbiamo la possibilità di leggere le recensioni di un luogo su internet oppure di sponsorizzare e diffondere le novità tramite i social.

È una nuova era quella per il turismo oggi che continua ad evolversi e ad adattarsi alle nuove tecnologie emergenti, potremmo tuttavia definire questa nuova forma di turismo come "Turismo 2.0".

1.3.1 Turismo legato ai social media

Negli ultimi anni i Social network hanno influenzato profondamente il nostro modo di comunicare con gli altri. Era impossibile che questo non interessasse anche un mondo dinamico e così legato alle opinioni delle persone come quello del turismo. La rivoluzione è stata, più che tecnologica, principalmente sociale: gli utenti (viaggiatori e operatori turistici) sono sempre più indipendenti, sono in grado di produrre contenuti e soprattutto di condividerli, di scambiarsi informazioni.

Tutto questo ha portato alla nascita di una nuova tipologia di viaggiatori, che si informa online senza bisogno di intermediazioni o di qualcuno che confezioni loro una vacanza pronta, vogliono essere loro stessi a cercare e a trovare ciò che desiderano.

Le comunità online e i social network caratterizzano oggi fortemente la scena del web e del marketing turistico, influenzando sempre di più le scelte dei viaggiatori. Rispetto al passato è cambiato molto: le modalità di accesso alle informazioni non sono più le stesse, sono aumentate le possibilità per chi vuole prenotare direttamente, per non parlare poi della nuova percezione della credibilità e della reputazione di destinazioni e di strutture che ormai dipende pesantemente dai commenti o dalle recensioni che i viaggiatori lasciano sui blog o sui social (es. Trip Advisor o Booking.com).

1.3.2 Turismo legato al Semantic Web

Quando si parla di web semantico si intende proporre un web che possieda delle strutture di collegamenti più espressive di quelle attuali. Il termine ‘Semantic Web’ è stato proposto per la prima volta nel 2001 da Tim Berners Lee. Da allora il termine è stato associato all’idea di un web nel quale agiscano agenti intelligenti: applicazioni in grado di comprendere il significato dei testi presenti sulla rete e perciò in grado di guidare l’utente direttamente verso l’informazione ricercata, oppure di sostituirsi a lui nello svolgimento di alcune operazioni. Un agente intelligente dovrebbe essere una applicazione in grado di svolgere operazioni come ad esempio la prenotazione di un aereo per Parigi con arrivo in centro città prima delle 13.00. Il tutto spulciando informazioni da siti che definiscono l’aeroporto di Parigi in modo diverso (Paris,

Charles de Gaule, Orly) e deducendo, senza che sia specificato nella query, che un arrivo per le 13.00 in centro implichi un arrivo in aeroporto diverso a seconda dell'aeroporto effettivamente selezionato. L'idea di Tim Berners Lee consisteva in una migrazione dal Web dei documenti al Web dei dati. Lo scopo del Web di dati è quello di collegare i concetti e contenuti reciprocamente, invece di collegare solamente documenti.

Così il Web dei dati ha portato alla conversione di documenti esistenti in dati collegati (Linked Data [5,2]) e alla creazione di nuovi insiemi di dati. Tra questi, uno dei gruppi di dati più sfruttato, che è stato il motore dell'intero progetto, è DBpedia[1], ovvero la versione in Linked Data di Wikipedia[2].

DBpedia [1] è disponibile in diverse lingue; la sua versione inglese contiene circa 4,0 milioni di cose, classificate in diverse categorie, tra cui persone, luoghi, lavori creativi, organizzazioni, specie e malattie.

Tuttavia, DBpedia, come Wikipedia, contiene solo un piccolo numero di informazioni relative al dominio del turismo, come ad esempio diversi alloggi e ristoranti. Inoltre, per quanto ne sappiamo, esistono solo pochi gruppi di linked data nel campo del turismo. Tra questi, bisogna citare El Viajero4, che fornisce informazioni con più di 20.000 guide di viaggi, foto, video e post, e anche di alloggi in Toscana con l'elenco delle varie strutture.

---

1 http://it.dbpedia.org/

2 http://wikipedia.or

1.4 Introduzione a Tourpedia

Nel mio elaborato di tesi, verrà illustrato Tourpedia[3], che potremmo definire il "Dbpedia del turismo". Tourpedia [7] è disponibile all'indirizzo oppure tramite datahub.io all'indirizzo ed è stato sviluppato all'interno del progetto OpeNER[4] (Open Polarity Enhanced Name Entity Recognition), il cui principale obiettivo è quello di attuare un sistema per l'elaborazione del linguaggio naturale. L'utilizzo di Tourpedia potrebbe essere molto diverso. Ad esempio potrebbe essere usato per effettuare una disambiguazione di un'entità nel settore del turismo o per estrarre il maggior numero di punti d'attrazione o di interesse in una città. Conoscere le entità associate ad ogni albergo è un passo fondamentale per il corretto funzionamento di Tourpedia.

Infatti tramite le entità riusciamo a trovare le attrazioni o i POIs nelle vicinanze di un albergo. Quindi per un risultato migliore è stato necessario perfezionare il sistema di estrazione delle entità (NERD) in modo che ad ogni albergo corrispondessero il maggior numero di entità.

Il mio lavoro consisteva appunto nell'estensione di un'applicazione precedente. È stata rinnovata sia la grafica dell'applicazione web che l'intero motore interno che gestisce i dati.

È stata aggiunta una sezione che contiene tutte le entità estratte e un'altra che contiene quelle filtrate per distanza e similarità. Infine sono state collegate tutte le entità e gli alberghi che si riferivano ad uno stesso posto e le rimanenti entità sono state aggiunte all'insieme degli alberghi.

---

3 http://tour-pedia.org

4 http://www.opener-project.eu/

1.5 Problema che ci poniamo

Il problema sorge quando un turista, in visita in una città, non sa esattamente dove pernottare e cosa poter visitare. Spesso accade che visitiamo delle città e non ci accorgiamo dei posti che possiamo visitare. Ad esempio un turista in visita a Firenze deve poter sapere che nella città potrà visitare tantissimi posti artistici e rinascimentali come il la Galleria degli Uffizi, il Duomo di Michelangelo e così via. Tourpedia cerca di risolvere questo problema andando a collocare i punti di interesse o le attrazioni più vicine agli alberghi (dove solitamente pernottano i turisti) in modo da offrire in servizio utile e funzionale. Inoltre è possibile creare vari itinerari sulla mappa così sarà impossibile perdersi

2 Lo stato dell'arte

Possiamo suddividere lo stato dell'arte è suddiviso in due sezioni, una parte che riguarda l'analisi di dataset simili, mentre una seconda parte che va più nello specifico e analizza strumenti per l'estrazione di Linked Data in modo simile al mio lavoro.

2.1 Analisi di dataset simili

Negli ultimi anni, molte iniziative sono state proposte nel campo dei dataset per il turismo. Qui diamo una panoramica unica degli insiemi di dati più importanti relativi al settore del turismo.

2.1.1 GeoNames

GeoNames è un dataset geografico molto particolare, con oltre 10 milioni di nomi geografici. Esso contiene anche luoghi turistici, come alberghi e ristoranti. Tuttavia, non fornisce alcun servizio SPARQL per interrogare i dati e, per ciascun nodo, dà solo poche informazioni, ad esempio, per l'Hotel Bologna a Pisa, esso fornisce solo le sue coordinate geografiche.

2.1.2 LinkedGeoData

È un'iniziativa aperta che esporta le informazioni estratte da OpenStreetMap come Resource Description Framework (RDF). LinkedGeoData contiene più di un miliardo di nodi. Dal momento che è derivato da una piattaforma di collaborazione, è in continuo aggiornamento. Per quanto riguarda LinkedGeoData, proponiamo una piccola base di dati, che contiene informazioni dettagliate per ciascun nodo. Per esempio, l'Hotel Bologna di Pisa contiene più informazioni in Tourpedia che in LinkedGeoData. In ogni caso, potrebbe risultare molto interessante per unire due basi di dati tramite una serie di collegamenti "sameAs" in modo da avere il maggior numero di informazioni disponibili per un albergo.

2.1.3 Salzburgeland Tourismus

Il dataset Salzburgerland Tourismus contiene tutti i dati rilevanti relativi a eventi, alberghi, locali e altro in Salzburgerland, Austria. Il dataset completo è disponibile all'indirizzo http://data.salzburgerland.com/dataset

2.2 Software per l'estrazione di Linked data

2.2.1 LODifier

LODifier[5] ha un approccio che coniuga un'approfondita analisi semantica con il riconoscimento del nome dell'entità e la disambiguazione del significato delle parole. Esso è composto da un vocabolario semantico al fine di estrarre l'entità con il nome associato e le relazioni con il testo convertendoli infine nella rappresentazione RDF legata a DBpedia e Word Net.

---

5 http://www.aifb.kit.edu/web/LODifier

2.2.2 NoDose

NoDose[6] è uno strumento interattivo per la semi estrazione di dati strutturati e semi strutturati da un file di testo. Utilizzando una GUI, l'utente decompone gerarchicamente il file, che delinea le sue regioni interessanti per poi descriverne la loro semantica.

Questo compito è accelerato da un componente che tenta di dedurre la grammatica del testo dalle informazioni che l'utente ha immesso finora. Una volta che il formato del documento è stato determinato, i dati possono essere estratti in un certo numero di forme utili. Questo documento descrive sia l'architettura, che può essere usata come banco di prova per algoritmi strutturali generali, e gli algoritmi sviluppati dall'autore.

3 Tourpedia

Tourpedia è un dataset collegato e strutturato di luoghi turistici estratti da quattro social media: Facebook[7], Foursquare[8], Google Places[9] e Booking.com[10].

Tourpedia contiene più di 500.000 places (luoghi) in Europa divise in quattro categorie: accomodation, restaurants, POIs e attractions. Per accomodation si intendono i luoghi dove si può dormire o stare in una location (hotels, B&B, etc....).

---

6 http://dl.acm.org/citation.cfm?id=276330

7 https://it-it.facebook.com/

8 https://it.foursquare.com/

9 https://plus.google.com/u/0/local.

10 http://www.booking.com/index.it.html

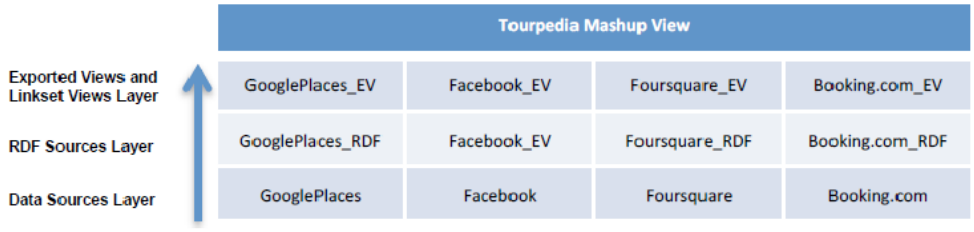

Fig. 2 Quattro livelli di frame work basati su ontologie

Un ristorante è inteso come un luogo dove si può mangiare o bere mentre per POI si intende un servizio locale per turisti come bancomat o biblioteche.

Infine per attraction indichiamo un luogo con attrazioni turistiche come musei o monumenti. In questo momento i paesi analizzati sono: Amsterdam, Barcellona, Berlino, Dubai, Londra, Parigi, Roma e la Toscana. Sono stati scelti questi luoghi per la loro popolarità in Europa anche se ci sono delle differenze, infatti ad esempio il dataset di Amsterdam è più piccolo rispetto a quello della Toscana.

Per costruire il set di dati, sono state usate le RESTfull API fornite dai summenzionati social media recuperando due differenti classi di informazioni che sono: informazioni richieste e informazioni opzionali. Le informazioni richieste includono il nome, la latitudine e la longitudine mentre le informazioni opzionali includono campi addizionali, tratti dai social media, come ad esempio gli indirizzi postali, numeri di telefono e descrizioni in varie lingue. Dato che i dati provengono da diverse fonti è stato necessario raggruppare i luoghi che rappresentano un unico oggetto reale in un'unica entità. Per questo è stato implementato un algoritmo di matching basato sulla similarità della distanza e sulla similarità delle stringhe. Come risultato ogni entità è rappresentata da un'unica risorsa disponibile tramite una URI http in accordo con il principio di Linked Data.

Con il termine Linked Data si fa riferimento ad una serie di pratiche per pubblicare e connettere tra loro attraverso il web collezioni di dati strutturati.

Negli ultimi anni la crescente richiesta da parte degli utenti di servizi che consentano un elevato grado di interconnessione tra i dati provenienti da diverse fonti ha portato all'adesione da parte di diversi provider di questa filosofia, con l'obiettivo di

costituire uno spazio dati globale e globalmente accessibile: il web dei dati, che costituisce la parte centrale del cosiddetto web semantico.

In quest'ottica, sono tre i fattori di maggiore interesse: la semantica, la fruibilità dei dati e la loro libera disponibilità per l'utente.

Libera disponibilità: i dati proposti sono Open, ovvero rilasciati secondo licenza Creative Commons - Attribuzione - Condividi allo stesso modo e liberamente utilizzabili dagli utenti;

Semantica: l'utilizzo di descrittori semantici consente di far emergere i collegamenti tra le risorse, correlandole in base al loro "significato". RDF fornisce in tal senso la possibilità di descrivere risorse tramite un modello dei dati strutturato ed orientato ai grafi, e costituisce pertanto il principale modello di descrizione utilizzato;

Fruibilità: all'interno della rete ipermediale i nodi divengono risorse, identificabili tramite uri e rappresentabili in vari formati, secondo le esigenze (RDF può ad esempio essere serializzato in formati xml così come su semplici file di testo, e presentato all'utente tramite delle pagine html)

3.1 L'importanza della posizione dell'albergo

Nessuna ricerca ha esaminato l'impatto della posizione dell'hotel e il successivo comportamento dei turisti. Intuitivamente, però, la posizione dell'hotel dovrebbe avere un profondo impatto sui movimenti turistici.

Arbel e Pizam (1977) hanno sostenuto quasi 40 anni fa, che la maggior parte dei turisti volevano trovarsi a pochi passi dalle principali attrazioni e quindi ricercavano gli alberghi in loro prossimità.

Più di recente, McKercher e Lau (2008), hanno scoperto che il 21% di tutte le escursioni fatte ad Hong Kong da turisti, si trovano a circa 500 metri dal loro albergo.

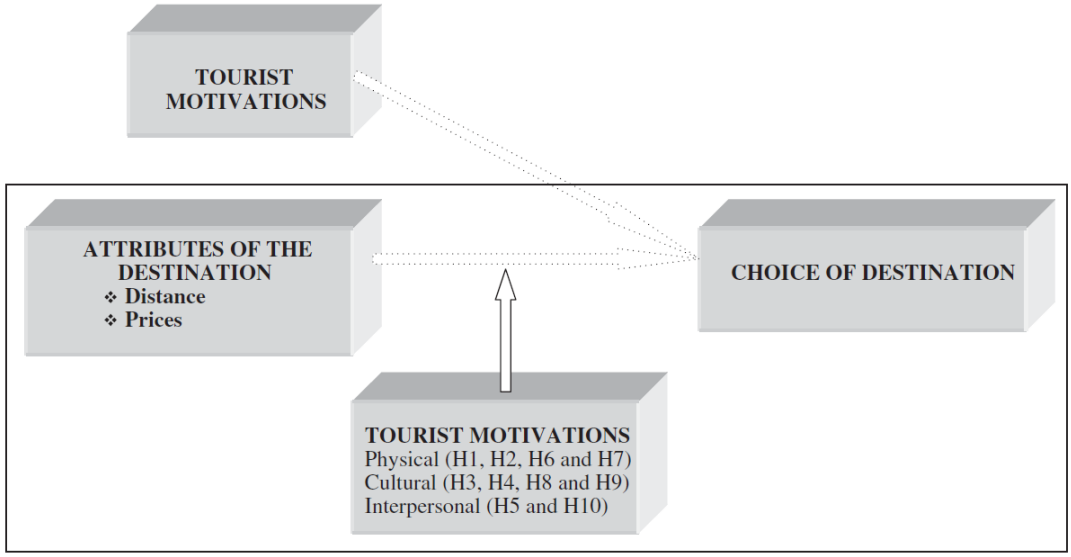

Fig. 4 Motivazioni turistiche per la scelta di una destinazione

Secondo una ricerca del "Department de Economı´a Financiera, Contabilidad y Marketing, Facultad de Econo´micas y Empresariales, University of Alicante, Ap. Correos 99, E-03080 Alicante, Spain" i fattori principali per la scelta di un albergo sono principalmente il fattore distanza e il fattore prezzo. Non è semplice capire con che criterio un turista sceglie una meta, vi sono tanti fattori da prendere in considerazione come il clima o la voglia di visitare il maggior numero di posti.

4 Estrazione di nuove entità

Nel corso degli ultimi anni, sono state implementate due categorie di ambienti di lavoro per cercare nuovi collegamenti, il primo consiste nell'antology matching [11, 9] mentre il secondo nell'istance matching [13].

La prima categoria prova a stabilire i collegamenti tra link di ontologie diverse, mentre la seconda categoria prova a stabilire i collegamenti tra istanze contenute in due datasets. Entrami assumono che esistono o due ontologie oppure due dataset da comparare. Questo significa che in ogni dataset o antologia, le antologie sono già definite e la funzione principale consiste nel trovare nuove relazioni [10] tra di esse.

Un esempio è LODifier [3], uno strumento che estrae le entità (NEs) e le loro relazioni con i testi e li esporta come Linked Data. Il tool usa le principali funzioni di NLP come la tokenizzazione, la Named Entity Recognition, il parser, la lemmatizzazione e la Word Sense Disambiguator, in ordine, per estrarre le relazioni attraverso le entità e associarle alle ontologie. In Tourpedia si prova a cercare le relazioni tra entità e entità citate nei testi associati.

4.1 Limiti e opportunità di Tourpedia

Nonostante la numerosa quantità di alberghi analizzati, Tourpedia, come tanti servizi del genere, soffre di alcuni limiti.

Il primo limite è la quantità di entità disponibili, infatti nonostante la NER ha trovato tantissime entità, è stato necessario filtrare solamente quelle geolocalizzate. Inoltre è stato necessario estendere Tourpedia aggiungendo e riorganizzando i collegamenti manualmente per far sì che un'entità come ad esempio "Train Station" si riferisse a "Pisa Train Station". La carenza di entità è dovuta anche alla massiccia presenza di alberghi senza la descrizione infatti, l'assenza di essa non permette al sistema di trovare l'entità associate.

4.2 Entity Discovery Sistem

L'idea principale per la scoperta di entità è trovare un modello linguistico per l'estrazione dei loro nomi. Tuttavia, i metodi tradizionali richiedono oltre ad un grande lavoro manuale, molto tempo. L'obiettivo è quello di trovare un algoritmo in grado di estrarre automaticamente le entità da un qualsiasi testo, facendo un'attenta analisi linguistica in modo da trovare le entità più significative con una probabilità più alta. Il sistema deve essere in grado di trovare il maggior numero di entità facendo un'accurata disambiguazione andando inoltre a rimuovere l'entità sbagliate.

Si comincia estraendo dalla categoria "accommodation" contenuta in Tourpedia, le descrizioni in inglese estratte da Booking.com.

Questo è l'esempio per l'Hotel Bologna di Pisa:

"The elegant Hotel Bologna is set in the historic centre of Pisa, 5 minutes' walk from the train station. It offers free Wi-Fi and a shuttle to the airport. All rooms have minibar, and flat-screen TV with satellite channels. They feature floor-to-ceiling windows, wooden floors and independent heating. Breakfast at the Bologna is a varied buffet including fruit, cold meats, and fresh bread and cakes. It can be enjoyed on the courtyard terrace, or provided to take away for early check-outs. The bar offers free tastings of traditional Tuscan products every evening. The Campo dei Miracoli and the famous Leaning Tower are a 12-minute walk from the hotel. The highway that links Pisa with Florence and Livorno is 3km away. Hotel Rooms: 64."

Il testo contiene un gran numero di POIs e attractions situate vicino all'accomodation (sottolineate nell'esempio). Questo succede per la maggior parte delle descrizioni infatti la presenza di POIs o attractions nelle descrizioni suggerisce di creare nuove relazioni tra esse e l'accomodation.

Prima di estrarre tutte le POIs/attractions citate nelle descrizioni, si è implementato un'Entity Discovery Sistem (EDS). Il tool estrae le descrizioni in lingua inglese e si interfaccia a DBpedia Spotlight che le analizza facendo la Named Entity Recognition and Disambiguation (NERD), restituendo un oggetto contenente le entità (dotate di link a dbpedia) che sono solamente dei luoghi, quindi dotati di coordinate geografiche. Questo processo è stato perfezionato cambiando i parametri di default di DBpedia Spootlight adattandolo ai testi in lingua inglese, cercando di aumentare il numero massimo di entità scoperte (che comunque verranno filtrate in seguito calcolandone la distanza dall'albergo.

La figura 5 mostra l'architettura dell'EDS, usato da tourpedia per trovare nuove relazioni tra entità. L'EDS è composto da tre moduli principali: il NE Extractor che prende le descrizioni delle accomodation in inglese e associa l'URIs al servizio di DBpedia Spotlight, il NE Filter che controlla quali entità sono dei luoghi e le filtra per distanza e similarità e il NE Linking che collega le entità che si riferiscono ad una stessa risorsa.

Una volta estratte le entità è stato indispensabile un lavoro di controllo manuale. Questa fase consiste nel correggere i links a DBpedia sbagliati e le corrispondenze

tra entità e place, come ad esempio dire che l'entità "Central Station" si riferisce a "Pisa Central Station" e non ad altre stazioni ferroviarie.

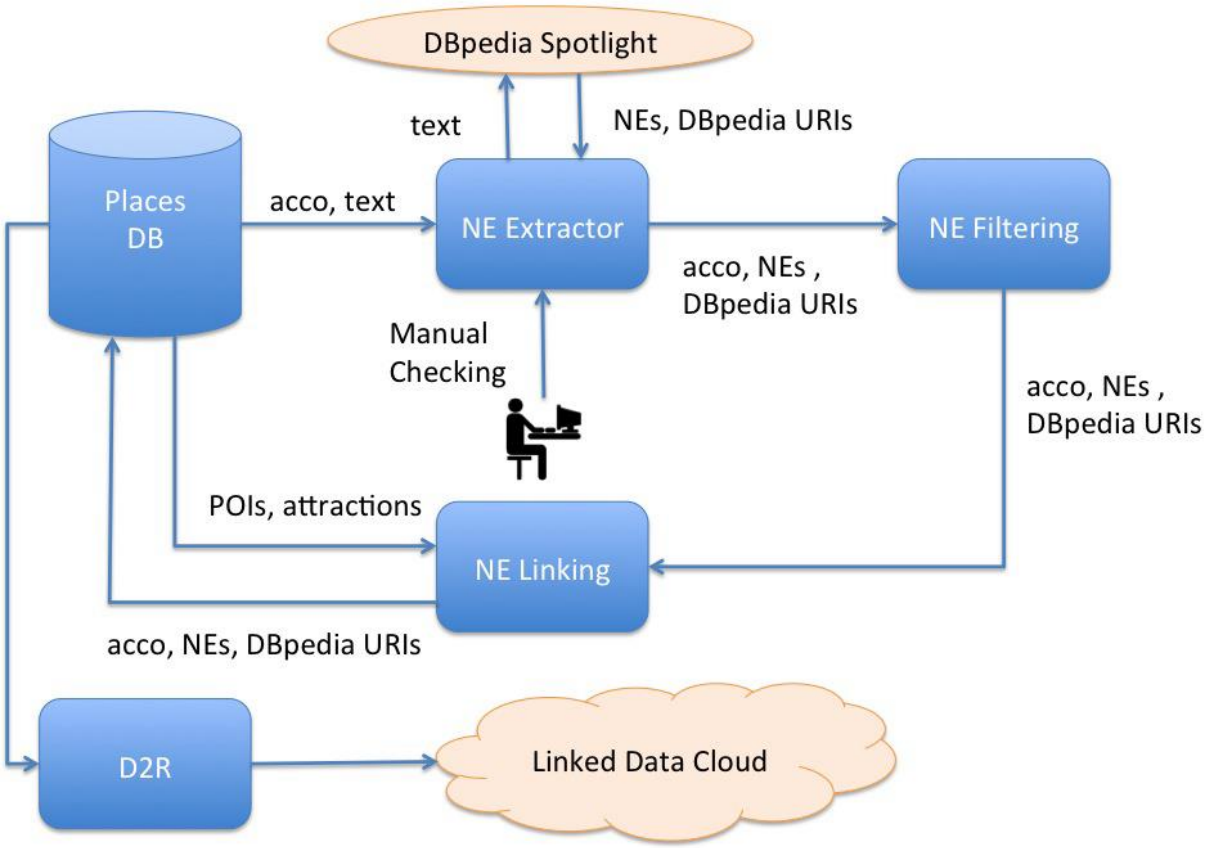

Fig. 5 L'architettura dell'EDS

Il modulo NE Filtering si occupa di andare a recuperare solamente le entità che si trovano a 10 chilometri tramite una funzione che calcola la distanza tra due posti e che sono dotate di latitudine, longitudine, descrizione e icona.

L'ultimo modulo è il NE Linking che collega le entità trovate con i place della tabella Places caratterizzate da proprietà simili. Questo significa che se sono riuscito a filtrare delle entità tramite NE Linking vuol dire che sono necessariamente dei luoghi!

4.2.1 NE Extractor

Il modulo NE Extractor è il modulo principale responsabile dell'estrazione delle entità. Esso esamina ogni descrizione associata ad ogni accomodation tramite le RESTful APIs provviste da DBpedia Spotlight. Per un'estrazione più efficiente, sono stati impostati diversi parametri nello script addetto al collegamento con le APIs tra essi ci sono ad esempio la confidenza impostata a 0.1 in modo da riuscire ad estrarre il maggior numero di entità (anche se non fossero state necessariamente di luoghi) oppure impostando la lingua in "en" in modo che DBpedia Spootlight capisse che si trattava di testi in lingua inglese. Lasciando tutti i parametri di default, venivano restituiti meno risultati, spesso errati.

```
class DBpediaSpotlight extends BaseAPI {
   public function init_nlp ($text) {
                $this->http_method = 'GET';
                $this->source_text = $text;
                $this->api_url = 'http://spotlight.dbpedia.org/rest/annotate';
                $this->api_args = array(
                        'text' => stripslashes($text),
                        'confidence' =>  0.1
                );
   }
```

Una volta estratte, le NEs sono state controllate manualmente, ovviamente questo controllo non è stato fatto per tutte le entità ma inizialmente per una piccola fetta in modo da valutare la qualità dei dati. In dettaglio il controllo consisteva nella disambiguazione di entità come "Train Station" o "Airport" delle quali non si conosceva in dettaglio a cosa ci si riferisse. L'obiettivo quindi è stato quello di andare a leggere la descrizione dalle quali sono state estratte ed andare a cercare e associare il link a Dbpedia corretto. Questa operazione è stata molto lunga sia perché le entità sono tante e poi perché bisogna andare correggere e soprattutto aggiungere nuove entità (creare una nuova risorsa, provista di "id" in modo da non sovrascrivere

o cancellare le altre presenti sul dataset). È in corso uno studio per rendere questo processo automatico o quanto meno semi-automatico.

4.2.2 NE Filtering

Il modulo NE_filter calcola la distanza tra albergo e entità ed è suddivisa in due parti: place recognition e place filtering.

La fase di place recognition filtra tutte le entità ottenute mantenendo solamente place, e quindi che sono dotate di coordinate geografiche. Questo processo sfrutta le informazioni ottenute dalla pagina di DBpedia associata a ciascun entità. Per ogni entità estratta NE, il modulo NE Filtering accede a DBpedia e controlla se l'entità ha le coordinate geografiche come proprietà (geo: lat and geo: long). Se ne è in possesso, essa è considerata un place, altrimenti viene scartata. È stata scritta una funzione che dato il link a DBpedia di qualsiasi entità controlla che si tratta effettivamente di un place e restituisce solamente quelle dotate di latitudine, longitudine, descrizione e icona.

```
function isPlace ($sDBpediaURL)
{
  $sSPARQLQuery = "PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
  PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
  SELECT ?lat ?long ?label ?description ?icon
  WHERE {
    {<" . $sDBpediaURL . "> dbpedia-owl:abstract ?description;
                foaf:depiction ?icon .
     FILTER(langMatches(lang(?description), 'EN'))}
    UNION
    {<" . $sDBpediaURL . "> geo:lat ?lat; geo:long ?long .}
    UNION
    {<" . $sDBpediaURL . "> a dbpedia-owl:Place; rdfs:label ?label .
    FILTER(langMatches(lang(?label), 'EN'))}
  }";
  $sBasicURL = "http://dbpedia.org/sparql?query=";
```

```
            $sURL = $sBasicURL . urlencode($sSPARQLQuery) . "&output=json";
            $ch = curl_init();
            curl_setopt($ch, CURLOPT_URL, $sURL);
            curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
            $aResult = json_decode(curl_exec($ch), true);
            $aReturn = array();
              If (isset($aResult['results']['bindings'][1]['long']) &&
isset($aResult['results']['bindings'][1]['long']['value']))
                $aReturn['long'] = $aResult['results']['bindings'][1]['long']['value'];
              else if (isset($aResult['results']['bindings'][1]['label']) &&
isset($aResult['results']['bindings'][1]['label']['value']))
          $aReturn['label'] = $aResult['results']['bindings'][1]['label']['value'];
              return $aReturn;
            }
            return false;
          }
```

La fase di place filtering scarta tutti i place che sono situati ad una distanza maggiore di 10 km (per convenzione) dall'accomodation facendo sì che la probabilità di associare un'accomodation con un'entità errata viene azzerata.

In dettaglio, per questa fase, si è usata la formula della "great circle distance".

Supponiamo di avere due punti A e B, con le rispettive coordinate (lat_a, $long_a$) e (lat_b, $long_b$), tutte espresse in radianti, la formula della great circle distance Δ in radianti tra A e B è data dalla seguente formula:

$$\Delta = acos(sin(lat_a)sin(lat_b) + cos(lat_a)cos(lat_b)cos(\Theta))$$

Dove $\Theta = abs|long_a - long_b|$. La distanza espressa in chilometri tra A e B è data da:

$$d(A.B) = r\Delta$$

Dove r è un'approssimazione del raggio della Terra, r = 6.371 Km.

La funzione che calcola la distanza è la seguente:

```
function calcola_distanza($latitude1, $longitude1, $latitude2, $longitude2)
{
$theta = $longitude1 - $longitude2;
$miles = (sin(deg2rad($latitude1)) * sin(deg2rad($latitude2))) + (cos(deg2rad($latitude1)) * cos(deg2rad($latitude2)) * cos(deg2rad($theta)));
$kilometers = $miles * 1.609344;
$meters = $kilometers * 1000;
return compact('meters');
}
```

4.2.3 NE Linking

Il modulo NE Linking stabilisce i collegamenti tra le entità ottenute NEs e i place contenuti nel DB Places, aventi come categoria attraction o POI. In particolare viene applicato un algoritmo di merging diviso in due fasi: distance similarity e string similarity. Durante la fase di distance similarity, solo le entità situate ad una certa distanza sono comparate.

In dettaglio, supponiamo che ε (ei $\subseteq$ ε) e inoltre, supponiamo che P è il contenitore di tutti i places con categoria attraction o POI e pk è un generico place in P (pk $\subseteq$ P). L'entità ei è comparata con tutti i pk che sono situati ad una distanza minore di una soglia 0 da ei (i.e. 10 km). Il primo algoritmo quindi riassume come la distance similarity viene calcolata ovvero, prende tutte le entità ei $\subseteq$ ε e pk $\subseteq$ P come input data la soglia 0. Come output viene prodotta la lista G di tutte le voci pk <= P|d (ei, pk) <= 0, $\subseteq$ei, dove d (ei, pk) si riferisce all'equazione 2.

L'algoritmo compila una nuova lista G[ei] con G, inizialmente vuota per ogni ei.

Quindi esso controlla tutte le entità pk e compara le distanze tra ei e pk. Se la distanza è minore o uguale alla soglia 0 allora l'entità viene aggiunta alla lista G[ei]

tramite la funzione push. Tutti gli accoppiamenti (ei, pk) sono processati durante la fase di string similarity, che esegue l'algoritmo di Similar Text per comparare le stringhe (i.e. nome entità/nome place).

```
        if (($distanza > 0) && ($distanza <= 10000)) {
            if ($similarità >20) {
                $match = "INSERT INTO ne_linking(id_entita,id_place,entita,place,distanza,db_link) VALUES ('$id_entita','$id_place','$nome_entita','$nome_place','$distanza','$db_link')";
                $query_match = mysql_query($match);
                if (!$query_match) {
                    die("Query non valida1: " . mysql_error());
                }
            }
        }
```

Alg. 1 Similarità tra stringhe e distanza

L'algoritmo 1 illustra la procedura per calcolare il matching della coppia (ei, pk). Esso prende G come input, e in base alla soglia impostata come percentuale di similarità (funzione "similar_text (x, y)") restituisce l'insieme M di tutte le coppie che hanno avuto un match (similarità >=20% e distanza >= 10 Km). In particolare, per ogni entità ei G viene inizialmente tokenizzata salvando tutti i tokens ti. Assumiamo che il nome dell'entità non contiene punteggiatura o simboli ma solamente il suo nome o il titolo. In questo modo la tokenizzazione è fatta solo tra due parole. Quindi l'algoritmo prende tutte le pk associate a ei (che sono contenute in G[ei]) e tokenizza le loro r-stringk. Adesso la procedura calcola la similarità tra le due stringhe come segue. Esso compara tutte le ti Tk tramite l'algoritmo di matching.

Se la similarità tra le due stringhe (tokens) è stata trovata (string_similarity (ti, tk)>= r), allora ss lo inserisco in NE_linking e passo il confronto alla prossima entità.

Quando tutti i tokens ti e tk sono stati comparati, ss viene normalizzata. Se ssn >= o le coppie (ei, pk) hanno avuto un match e sono state aggiunte in M.

La normalizzazione è fatta tramite la seguente formula:

$$ssn_2 \frac{ss}{+Nk}$$

Dove Ni e Nk rappresentano il numero di tokens in r-stringi e r-stringk, rispettivamente. Nota che 0 <= ssn <=1.

L'algoritmo di matching è una procedura che prende due tokens ta e tb come input e restituisce quanto sono simili come output. L'output esprime la tokens similarity espressa in percentuale. In dettaglio, si assume che ai è un carattere definito in entrambi (ta e tb) e Na e Nb sono i numeri dei caratteri in ta e tb, rispettivamente.

Quindi il valore di ts è calcolato come segue:

$$ts = 2 \frac{\sum_{i=i}^{min(Na,Nb)} ai}{Na + Nb}$$

Per migliorare le prestazioni dell'algoritmo, la lista delle stopwords potrebbe essere definita e un token appartenente a tale elenco dovrebbe essere escluso dal confronto. Ad esempio nel dominio delle accommodation, tutte le parole "Hotel", "Restaurant" etc. dovrebbero essere include nella lista delle stopwords.

4.3 Valutazione dei risultati

In questo paragrafo mostrerò i risultati ottenuti dalle varie operazioni e quanto il sistema è migliorato dopo di esse. Tutte le entità e proprietà contenute in Tourpedia sono esportate come Linked Data. Nella fase precedente [4], abbiamo descritto l'ontologia usata per rappresentare le accomodations contenute in Tourpedia. In più, l'output dell'EDS è usato per arricchire Tourpedia esportandolo come Linked Data.

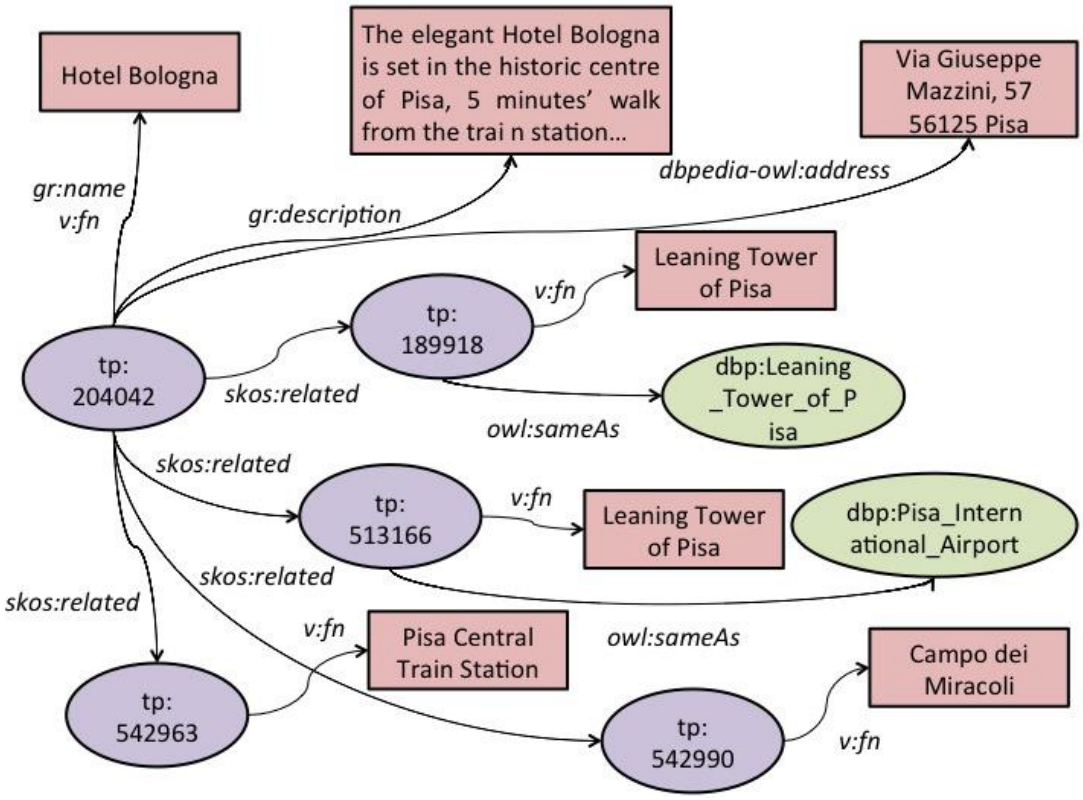

Fig. 2 Una risorsa Tourpedia arricchita con l'EDS

La figura 2 mostra un esempio di una risorsa di Tourpedia, i.e. l'Hotel Bologna a Pisa (Italy). La risorsa analizzata è identificata dall' ID 204042. L'EDS trova sette risorse in ordine per fare un esempio chiaro. Ogni risorsa scoperta è connessa all'id 204042 tramite la proprietà skos: related.

Le risorse trovate sono le seguenti: 1) 189918 (Leaning Tower of Pisa); 2) 513166 (Galileo Galilei Airport) e 3) 542963 (Pisa Central Station), che ho corretto durante la fase di Ne Extraction; 4) 542990 (Campo dei Miracoli), con l'aggiunta del dataset di Tourpedia, originariamente non contenuto in esso; 5) Pisa, 6) Livorno e 7) Florence, che non sono state aggiunte a Tourpedia, perché non rappresentano accommodations, restaurants, POIs o attractions. Le risorse 1), 2) e 3) sono connesse a DBpedia tramite la proprietà owl: sameAs.

Nella versione iniziale del progetto, erano contenuti sul database 525239 alberghi distribuiti in tutta Europa.

Per ogni albergo sono state analizzate le descrizioni (NER), estraendo qualsiasi tipo di entità dai testi ottenendo un totale di 9814 entità. In una nuova tabella sono state salvate tutte le coppie hotel-entità raggiungendo un totale di 88314 coppie diverse. Successivamente il modulo NE_filter ha calcolato la distanza di tutte le coppie hotel-entità ottenendo 16049 coppie di luoghi e successivamente ha inserito in una nuova tabella (NE_filter) solamente le coppie con distanza massima di 10 chilometri che sono esattamente 1064.

In seguito tramite la NE_linking sono state connesse 809 entità tutte dotate del link a DBpedia o a Wikipedia (in alcune entità sistemate manualmente).

L'entità che non sono state connesse (EP) sono state aggiunte alla tabella places tramite la seguente formula:

$$EP = NE\_filter - NE\_linking$$

Quindi all'insieme degli alberghi sono state aggiunte 255 entità.

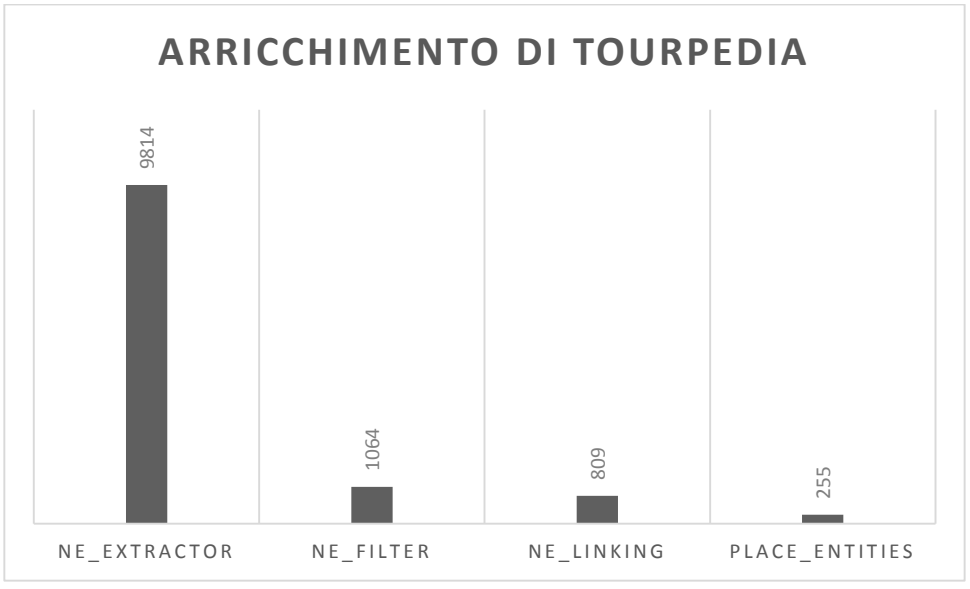

Fig. 6 Arricchimento di Tourpedia

5 L’applicazione web

Nel precedente lavoro, abbiamo implementato un’applicazione web, che gestisce i contenuti di Tourpedia. Lo abbiamo esteso in modo da supportare le relazioni tra le entità trovate (figura 3).

In particolare l’applicazione web adesso permette all’utente di visualizzare sulla mappa le entità di ogni albergo, leggere le descrizioni associate, visualizzare la foto del posto, creare l’itinerario più breve tra l’accommodation e le entità, e visualizzare tutte le caratteristiche di un albergo.

Il lavoro che ho svolto si basa sull’estensione di un’applicazione esistente.

5.1 Versione base

La versione precedente di Tourpedia era molto diversa graficamente rispetto alla nuova versione ed è possibile vederla nella figura 7. Il sito si presentava con una grande mappa a tutto schermo e un semplice menù statico in cima alla pagina con il quale era possibile scegliere un paese. Inoltre era presente una barra di ricerca situata in altro a destra che permetteva di ricercare i vari alberghi nei paesi europei.

Come nella nuova interfaccia venivano mostrate le icone a forma di faccina sulla mappa che segnalavano la presenza di un albergo e cliccandoci sopra veniva mostrata una sezione dove era possibile vedere la locazione dell’albergo e le reviews.

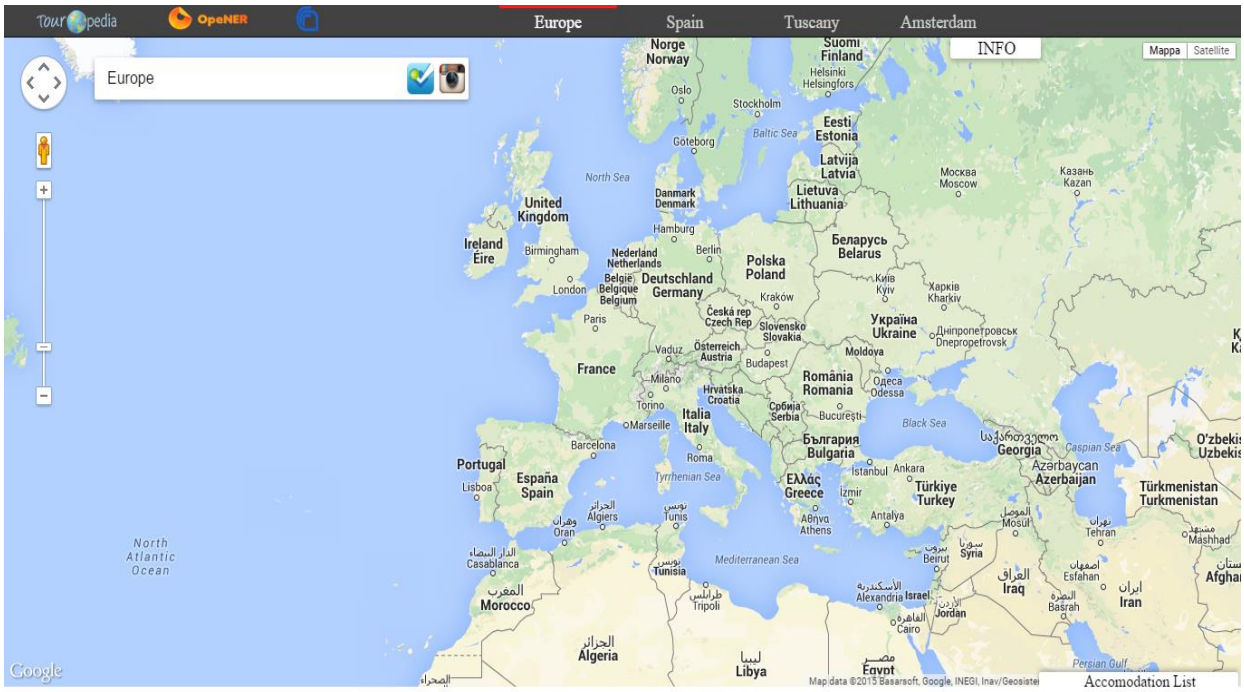

Fig. 7 Vecchia Interfaccia

5.2 Versione estesa

Rispetto alla precedente versione, sono state effettuare diverse modifiche sia per l'aspetto estetico dell'applicazione, sia per le funzionalità. La figura 8 mostra un esempio della nuova veste grafica di Tourpedia.

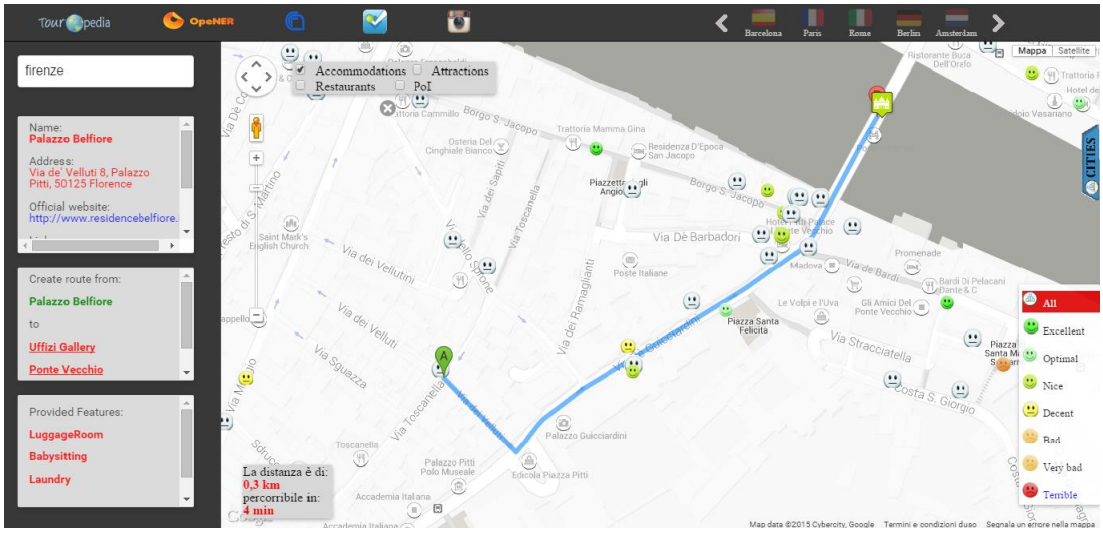

Fig. 8 Nuova interfaccia

Si può notare che la grafica in generale ha subito un netto miglioramento. È stato cambiato il tema del sito scegliendo dei colori più scuri sia per il menù e la mappa che per la barra laterale che è stata aggiunta.

Sulla barra superiore è disponibile un menu di navigazione con il quale è possibile scorrere diversi paesi europei, oscurati se non attivi (Toscana, Amsterdam, Barcellona, Dubai, Berlino, Roma, Parigi, Londra), tramite due freccette e sceglierne uno. In alternativa è possibile ricercare il paese o la città tramite la barra di ricerca situata in alto a destra sotto il logo di Tourpedia. Una volta che viene cercata una città verranno visualizzate sulla mappa tutte le accommodations (faccine di colore diverso che vanno dal verde al rosso), della città. Quando l'utente clicca su un albergo, vengono create 3 nuove sezioni nella pagina situate sulla side bar di sinistra.

La prima sezione contiene le informazioni relative all'albergo come il nome, l'indirizzo, il sito web, il link a Google places e le recensioni lasciate dai viaggiatori.

La seconda sezione mostra l'entità associate all'albergo di colore rosso sottolineato e permette di visualizzare l'itinerario sulla mappa. Qualora l'utente decidesse di cambiare destinazione, cliccando su una nuova entità, verrà cancellato dalla mappa il vecchio percorso e ne verrà creato uno nuovo.

La terza sezione invece mostra le caratteristiche disponibili associate ad un albergo. Un esempio potrebbe essere la disponibilità di connessione Wi-Fi, l'aria condizionata, parcheggio auto e così via. Questa sezione potrebbe risultare molto importate per quei turisti che hanno dei bisogni particolari e hanno intenzione di sapere se l'albergo potrà soddisfare i loro bisogni.

Tourpedia inoltre mostra un'analisi sentimentale degli alberghi su una mappa, che in questo caso è Google Maps[11].

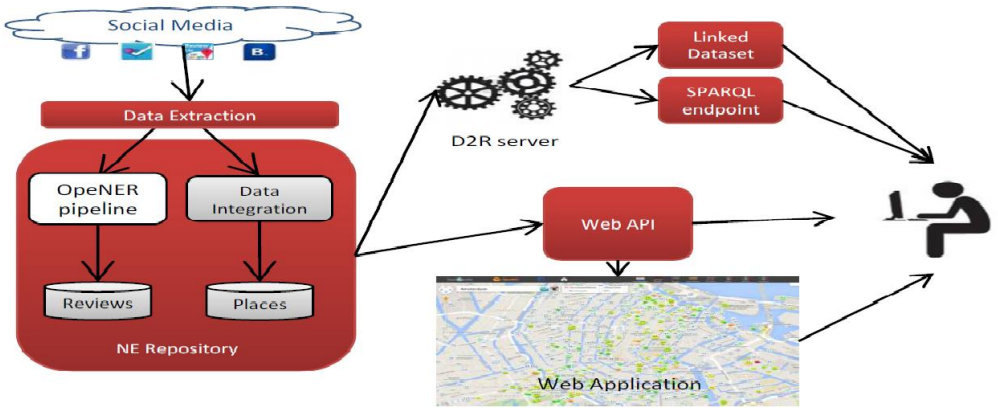

Fig. 9 Estrazione di dati e visualizzazione

L'analisi sentimentale di un luogo è calcolata in funzione di tutte le recensioni su quel luogo. Per recuperare l'analisi sentimentale di un luogo tramite OpeNER, viene fatta un'analisi particolare che si basa sui commenti (reviews) delle persone. Ogni commento viene elaborato da OpeNER al quale viene associato un tasso, che sarà la metrica di valutazione del sentimento. In base al sentimento varia l'immagine a forma di faccina che verrà visualizzata sulla mappa. Inoltre è stato aggiunto alla mappa un nuovo menù a scomparsa che permette di visualizzare sulla mappa solamente gli alberghi con una certa polarità.

Adesso è possibile scegliere se visualizzare gli alberghi considerati eccellenti, ottimi, discreti, pessimi e terribili

11 https://www.google.it/maps

5.3 Aggiunta di entità

La novità di Tourpedia consiste nel fatto di visualizzare le entità di un albergo nell'arco di 10 chilometri. Esse vengono visualizzate quando un turista clicca su un marker di albergo visualizzato sulla mappa (le faccine colorate). Cliccandoci sopra infatti appariranno sulla mappa tutte le entità associate ad un albergo, con un effetto stile pioggia, distinguibili dai marker di colore giallo.

In più, vengono aggiunte 2 sezioni in basso a sinistra della pagina che sono:

Sezione in cui vengono mostrate le entità associate ad ogni albergo che viene cliccato con la possibilità di creare un percorso dall'hotel all'entità cliccata

Sezione in cui vengono visualizzate tutte le caratteristiche fornite da un albergo come ad esempio il Wi-Fi, l'aria condizionata, parcheggio auto ecc.…

5.4 Creazione percorsi

Quando vengono generate le entità sulla mappa, viene creata una sezione relativa alla generazione di un percorso dall'albergo all'entità. Le entità sono distinguibili da un font di colore rosso sottolineato, per far sì che l'utente capisce che si tratta di link cliccabili. Quando viene selezionata un'entità, verrà generato un percorso sulla mappa che condurrà il turista alla destinazione scelta. Se dovesse scegliere di cambiare meta, verrà cancellato il vecchio percorso e ne verrà creato uno nuovo

Conclusioni e lavori futuri

Nella mio elaborato ho descritto Tourpedia, un dataset geografico connesso, derivato dai dati estratti dai social media. In particolare, ho descritto come Tourpedia trova e aggiunge le nuove relazioni tra entità all'intero del dataset mediante l'uso di un Entity Discovery System (EDS). Al momento l'EDS estrae solo le entità dai testi in lingua inglese. Potrebbe essere interessante analizzare testi in altre lingue dato che Booking.com ne è provvisto per avere le traduzioni dei nomi delle entità. Come lavoro futuro è previsto di rilasciare un frame work pubblico per l'EDS, in modo da analizzare qualsiasi dataset con il suddetto tool.

Ringraziamenti

Volevo ringraziare innanzitutto i miei genitori che mi hanno permesso di affrontare questo percorso di studi serenamente e che mi hanno sempre sostenuto nonostante le tante difficoltà del mio percorso. Un altro ringraziamento va a tutti i miei compagni di studio che mi hanno sempre aiutato durante gli esami più difficili, che mi hanno accompagnato fino alla fine di questo stupendo percorso di studio.

Inoltre volevo ringraziare i miei relatori Andrea Marchetti e in particolare Angelica Lo Duca per avermi guidato in questo progetto dedicandomi buona parte del suo tempo aiutandomi e rispondendo ai miei dubbi.

Bibliografia

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika,D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudr~ Al'-Mauroux, editors, The Semantic Web, volume 4825 of Lecture Notes in Computer Science, pages 722{735. Springer Berlin Heidelberg, 2007.

[2] S. Auer, J. Lehmann, and A.-C. Ndonga Ngomo. Introduction to linked data and its lifecycle on the
web. In A. Polleres, C. d'Amato, M. Arenas, S. Handschuh, P. Kroner, S. Ossowski, and P. Patel-Schneider, editors, Reasoning Web. Semantic Technologies for the Web of Data, volume 6848 of Lecture Notes in Computer Science, pages 1{75. Springer Berlin Heidelberg, 2011.

[3] I. Augenstein, S. Pad_o, and S. Rudolph. Lodi_er: Generating linked data from unstructured text. In
Proceedings of the 9th International Conference on The Semantic Web: Research and Applications, ESWC'12, pages 210{224, Berlin, Heidelberg, 2012. Springer-Verlag.

[4] C. Bacciu, A. Lo Duca, A. Marchetti, and M. Tesconi. Accommodations in Tuscany as Linked Data. In Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC 2014),
pages 3542{3545, May, 26-31 2014.

[5] C. Bizer, T. Heath, and T. Berners-Lee. Linked data -the story so far. International Journal on Semantic Web and Information Systems, 5(3):1^a_A _S22, 2009.

[6] S. Cresci, A. D'Errico, D. Gazzè, A. Lo Duca, A. Marchetti, and M. Tesconi. Tourpedia: a web application for sentiment visualization in tourism domain. In Proceedings of The OpeNER Workshop in LREC 2014, pages 18{21, 2014.

[7] S. Cresci, A. D'Errico, D. Gazzè, A. Lo Duca,A. Marchetti, and M. Tesconi. Towards a DBpedia of Tourism: the case of Tourpedia. In Proceedings of the 2014 International Conference on Semantic Web - Poster and Demo Track, ISWC2014, pages 129{132, 2014}.

[8] K. Gade. A non-singular horizontal positionrepresentation. Journal of Navigation, 63:395{417, 7 2010.

[9] Y. Jean-Mary, E. Shironoshita, and M. Kabuka.Ontology matching with semantic veri_cation. Web Semantics: Science, Services and Agents on the World Wide Web, 7(3), 2009.

[10] A. Koukourikos, G. A. Vouros, and V. Karkaletsis. Towards enriching linked open data via open
information extraction. In _rst international workshop on Knowledge Discovery and Data Mining Meets
Linked Open Data (Know@LOD), 2012.

[11] P. Lambrix and H. Tan. Sambo a system for aligning and merging biomedical ontologies. Web Semantics: Science, Services and Agents on the World Wide Web, 4(3):196 { 206, 2006. Semantic Web for Life Sciences.

[12] I. Oliver. Programming Classics: Implementing the World's Best Algorithms. Prentice Hall, 1993.

[13] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In Proceedings of the 8th International Semantic Web Conference, ISWC '09, pages 650{665, Berlin,
Heidelberg, 2009. Springer-Verlag.

Sitografia

- OpeNER, The OpeNER project http://www.opener-project.eu/ (visitato il 19 Maggio 2015).
- Booking.com , http://www.booking.com (visitato il 19 Maggio 2015)
- DBpedia Spootlight, https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki (visitato il 19 Maggio 2015)
- Department de Economı´a Financiera, Spain, http://www.journals.elsevier.com/tourism-management (visitato il 19 Maggio 2015)

Riconoscimenti

Questo lavoro è stato realizzato nell'ambito del progetto opener, co- finanziato dalla Commissione europea sotto FP7 (7th Framework Programs Grant Agreement n. 296451).