UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea Triennale in Ingegneria Informatica

Geotagging di tweet mediante annotatori semantici

Tesi di Laurea

Candidato:
Gabriele Serra

Relatori:
Prof. Marco Avvenuti
Prof. Maurizio Tesconi
Ing. Stefano Cresci
Ing. Fabio Del Vigna

Anno Accademico 2015/2016

Sommario

L'utilizzo quotidiano di social media e la condivisione di informazioni real-time da parte degli utenti può essere sfruttata per situazioni di emergenza sociale e disastri come in caso di sisma. Gli utenti da questo punto di vista vengono considerati dei *social sensors* che forniscono informazioni sempre aggiornate sul luogo in cui si trovano.

In questo documento sarà descritta la progettazione e l'implementazione di un sistema per la localizzazione spaziale di messaggi provenienti da Twitter (uno dei social media più utilizzati) detti *tweets*. L'analisi è effettuata tramite l'utilizzo di strumenti di analisi linguistica, gli annotatori semantici.

I tweets dopo essere stati salvati in un database relazionale vengono analizzati dal sistema che cerca all'interno di ogni tweet un luogo geografico. Una volta trovato vengono ottenute le coordinate del luogo e salvate all'interno del database.

Con questo approccio sono stati analizzati dati di alcuni fenomeni accaduti in Italia e non. Le risposte degli strumenti di analisi sono state valutate tramite l'impiego di matrici di confusione e inoltre sono state create delle mappe di crisi di tipo coropletico in modo da poter visualizzare l'output. I risultati ottenuti sono molto soddisfacenti in quanto l'abilità degli annotatori nel riconoscere i luoghi è risultata molto alta e si avvicina alla certezza statistica, inoltre il numero cospicuo di tweets permette di costruire mappe accurate anche dopo solamente un'ora dall'evento disastroso.

In futuro il sistema sviluppato potrebbe essere migliorato e utilizzato anche con stream di dati real-time in modo da possedere uno strumento versatile che può davvero aiutare le autorità soprattutto nei soccorsi e nella raccolta dati nelle ore immediatamente successive al fenomeno.

Indice

Elenco delle figure

Elenco delle tabelle

1. Introduzione

I social network sono popolari mezzi di condivisione di informazioni tra reti di utenti. Di fatto ogni giorno milioni di utenti sono attratti dai più utilizzati social media che acquisiscono da essi informazioni riguardanti interessi personali, preferenze, attività ed eventi. I social media quindi offrono la possibilità di accedere a tutte queste informazioni e dati senza che gli utenti stessi debbano metterli forzatamente a disposizione, bensì come side-effect dell'utilizzo quotidiano [1]. Da parte di chi utilizza il social media infatti non è richiesto un impegno costante e in quest'ottica gli utenti possono essere considerati dei *social sensors* cioè fonti di informazioni in tempo reale e non, relativamente all'ambiente che li circonda e alle esperienze che stanno vivendo.

Uno dei settori più interessanti per l'applicazione di questo paradigma è quello della gestione di crisi ed emergenze. In caso di disastro naturale infatti può essere fondamentale collezionare in modo veloce ed efficiente informazioni fornite da utenti che si trovano nelle vicinanze del fenomeno [2]. Il contributo che i social media forniscono può essere determinante per salvare delle vite umane o rilevare zone colpite dal disastro e richieste di soccorso. Alcuni social media già adesso hanno implementato dei meccanismi per aiutare la collettività in caso di disastri naturali. Facebook, il social media più utilizzato nel mondo, con più di 1,6 miliardi di iscritti, mette già a disposizione delle funzioni di questo tipo. L'esempio più concreto è rappresentato dalla funzione *safety check* che permette di segnalare agli altri utenti della propria rete sociale il proprio stato di salute in una situazione d'emergenza.[1] L'elaborazione dei dati proveniente da un social media può essere ancora più approfondito. In questo documento ci soffermeremo sull'analisi di messaggi pubblicati dagli utenti su Twitter.

Di fatto Twitter[2] è un servizio di microblog che conta più di 500 milioni di iscritti e addirittura 300 milioni di utenti attivi giornalmente ed è particolarmente idoneo allo studio di dati di questo tipo. Gli utenti infatti possono condividere pubblicamente dei brevi messaggi

[1] È possibile saperne di più all'url: https://www.facebook.com/about/safetycheck/

[2] Le statistiche di utilizzo possono essere trovate qui: https://about.twitter.com/

detti *tweets*. Questi messaggi sono pubblicati in real-time e accessibili da chiunque. Alcuni dati estraibili dai tweets possono essere l'entità del danno a cose o persone oppure il luogo geografico dal quale il messaggio stesso proviene.

L'approccio seguito nel documento quindi prevede l'analisi di questi messaggi tramite strumenti linguistici detti *annotatori semantici* al fine di ottenere in modo automatico delle informazioni sempre aggiornate [3]. Gli strumenti detti annotatori semantici sono capaci di analizzare frasi o testi, capirne il significato (*topic*) e suddividere di conseguenza il testo stesso in sotto stringhe chiamate *spots* [4]. Ogni spot, se ha un significato pertinente o utile nel contesto del topic, viene collegato alla rispettiva pagina su un database di conoscenza (es. Wikipedia, DB Pedia[3]). L'associazione è quindi formata da una coppia [spot,entità della *knowledge-base*] a cui viene associato un valore detto confidenza che rispecchia la qualità dell'associazione stessa. L'annotatore infatti non può essere sicuro di aver collegato una voce correttamente. Inoltre in caso di omonimi sullo stesso database di conoscenza, viene fatta una disambiguazione che di solito dipende dal significato complessivo attribuito dall'annotatore al testo stesso nella sua interezza. L'annotazione di un testo di solito è organizzata in tre task differenti e consecutivi. Il primo consiste nell'individuare i frammenti di testo che possono essere riferiti ad una entità della base di conoscenza. Questi frammenti sono gli spot. Il secondo task consiste nella disambiguazione dell'associazione che si traduce nella scelta della corretta entità proveniente dalla base di conoscenza in caso di entità omonime ma con significati diversi fra loro. Il terzo e ultimo task è quello di escludere, tramite alcuni parametri, come la confidenza, le annotazioni non consone al topic del testo. Il processo di annotazione quindi va quindi oltre al semplice arricchimento di un testo con collegamenti, ed è legato alla contestualizzazione di ogni spot e all'attribuizione del corretto significato generale. Gli annotatori semantici utilizzati in questo documento sono Tagme [4], Dexter [5] e Spotlight [6].

In particolare ci focalizzeremo sull'utilizzo degli annotatori semantici per ricavare il luogo di provenienza dei vari tweets e per valutare la bontà dell'annotazione. Un processo di questo tipo è detto *geoparsing*. Il geoparsing consiste nell'estrarre da un *plain-text* un luogo geografico non ambiguo, riconducibile quindi ad una voce su un database di conoscenza oppure riferibile tramite delle coordinate geografiche.

L'importanza di questo task è fondamentale, infatti alcuni studi evidenziano che i tweets geo-taggati nativamente sono solamente una piccola percentuale del totale [7]. Il geoparsing di plain-text tramite annotatori semantici oltre ad essere estremamente flessibile garantisce la certezza di non commettere errori causati da polisemia toponomastica [3]. Infatti l'annotatore

[3]DBpedia è un progetto aperto e collaborativo per l'estrazione di informazioni semi-strutturate da Wikipedia, per approfondire: http://wiki.dbpedia.org/

semantico esegue automaticamente la disambiguazione. Di fondamentale importanza inoltre è la valutazione delle risposte fornite dai tre diversi annotatori.

La comparazione dell'operato di strumenti di annotazione non è ovvia dal momento che non esistono metriche specifiche, ma in generale ci si affida a metriche più tradizionali, adottate largamente nell'ambito del machine learning. La valutazione dei risultati infatti spesso è calcolata con metriche difformi tra loro e su dataset diversi. Il problema di comparare i risultati quindi non è intrinsecamente legato al task di annotazione. Si stima che ricercatori che lavorano in questo settore spendano fra il 60 e l'80% del loro tempo a preparare e comparare input e output per l'annotatore semantico [8].

2. Related Work

Esistono in letteratura numerosi dibattiti sull'utilizzo di annotatori semantici per l'analisi di informazioni provenienti da social media. Ovviamente anche la valutazione delle risposte è parte integrante del dibattito.

2.1 Geoparsing

2.1.1 Geoparse di Tweets con NLTK

Uno degli approcci più utilizzati in letteratura è quello del geoparse on-the-fly dei tweets usando strumenti di analisi di linguaggio naturale. Un approccio di questo tipo è stato utilizzato nella pubblicazione dell'University of Southampton [9]. In questo caso ogni tweet viene analizzato con NLTK (Natural Language ToolKit) [10]. L'approccio seguito consiste nel precaricare nel sistema le entità geografiche principali del luogo in cui è avvenuto il disastro. Successivamente viene analizzata la frase e suddivisa in token di lunghezza variabile escludendo le *stop words*.[1] La fase successiva consiste nel confrontare ogni token ottenuto con le entità geografiche contenute nella tabella precaricata nel sistema. In questo modo vi è una altissima probabilità di geotaggare ogni luogo all'interno del tweet. Questo tipo di approccio richiede però la conoscenza a priori della località in cui è avvenuto l'evento. Questo non consente di implementare uno strumento flessibile. Ogni qualvolta si desideri modificare la zona controllata è necessario ripetere il task di precarica delle entità geografiche sulla tabella. Inoltre un approccio puramente NLP[2] deve possedere dei modelli d'analisi molto avanzati che dipendono dalla lingua in cui è scritto il testo.

[1] È detta stop word ogni parola generica e di uso comune, come articoli e congiunzioni, che ogni lingua possiede.

[2] NLP acronimo che significa *Natural Language Processing*

2.1.2 Content-based geoparse

La soluzione proposta dalla Texas A&M University invece prevede di geolocalizzare il tweet tramite un approccio probabilistico basato sull'analisi del contenuto del testo [7]. L'intuizione del team che ha lavorato in questa direzione è semplice ma innovativa. Un tweet infatti può contenere alcuni termini specifici relativi al luogo in cui è stato scritto il tweet oppure locuzioni associabili a determinati luoghi (es. esclamazioni tipiche dialettali). Gli esperimenti proposti dal paper mostrano che un framework di questo tipo funziona molto bene se i tweets contengono all'interno alcuni riferimenti spaziali. In assenza di essi è molto difficile stimare il luogo d'origine del tweet. In entrambi i casi comunque emerge il limite di questo tipo d'approccio dato che la localizzazione non è precisa e di conseguenza non potrebbe essere utilizzata in un task di geotagging per situazioni di emergenza in cui si richiede la massima accuratezza.

2.2 Valutazione di task d'annotazione

2.2.1 Valutazione con Framework automatici

La valutazione in termini statistici di strumenti di annotazione ha portato allo sviluppo di framework automatici per la comparazione di annotatori come GERBIL [8].

L'approccio di tipo automatico però servire a valutare l'operato di un annotatore nel complesso e fornisce delle metriche di tipo statistico che difficilmente si adattano allo scopo che vogliamo raggiungere ovvero ottenere informazioni riguardo luoghi geografici con la possibilità di integrare l'output di un annotatore con altri tipi di informazioni già in nostro possesso.

3. Lavoro Svolto

3.1 System Overwiew

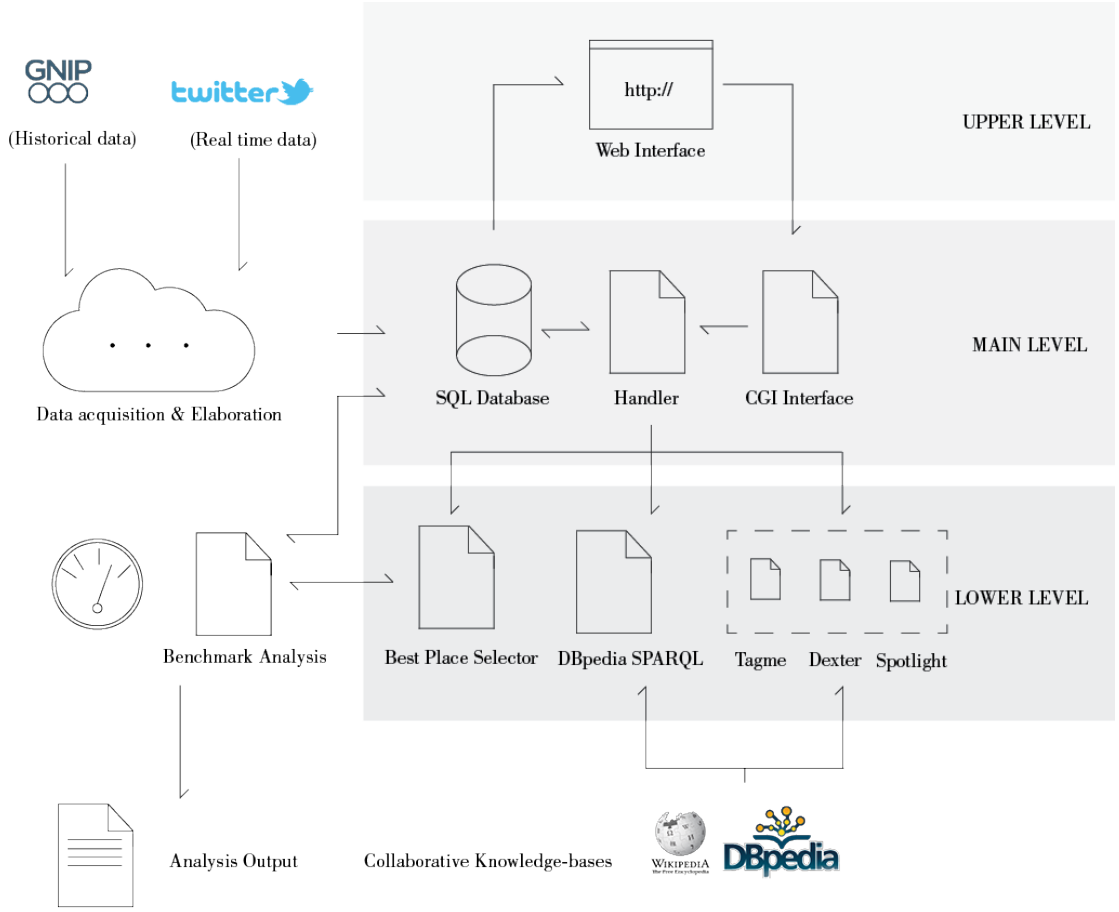


Figura 3.1: *Architecture overview del sistema sviluppato.*

Il sistema sviluppato permette tramite un interfaccia Web di analizzare un dataset di tweets scegliendo sia l'annotatore preferito sia le varie impostazioni riguardanti la lingua e il *tune*

(es. la confidenza minima richiesta). Inoltre le risposte fornite dai vari annotatori vengono ulteriormente elaborate in modo da ottenere le coordinate relative ad una entità geografica e di conseguenza scegliere l'entità migliore per ogni tweet. Il sistema è disponibile all'url: http://wafi.iit.cnr.it/crismap/crismap/geoparsing/. È possibile visionare il modello architetturale in figura. Il livello più alto fornisce l'interfaccia di tipo Web allo strumento in modo da poter utilizzare in maniera intuitiva il modulo a livello sottostante, impostando i parametri che servono allo strumento per poter operare. Naturalmente l'interfaccia fornisce anche un output dello strumento, organizzata in maniera tabellare. L'interfaccia grafica Web e il livello sottostante dialogano tramite Common Gateway Interface[1]. La tecnologia CGI permette l'esecuzione da parte del server di un applicazione esterna e, a tale riguardo, la tecnologia CGI non è altro che una forma di Remote Procedure Call[2]. In questo modo la pagina Web HTML richiede al server di eseguire in tempo reale il modulo a livello intermedio per generare dinamicamente le informazioni relative al task. La richiesta al server avviene tramite AJAX (Asynchronous JavaScript and XML) in modo da nascondere e rendere più naturale all'utente l'elaborazione da parte del server. Il livello intermedio richiede servizi forniti dai moduli a livello più basso, inoltre salva le varie annotazioni sul database da cui ha ricevuto i dati in un primo momento. Il livello intermedio svolge quindi sequenzialmente le varie azioni richieste per completare tutto il task d'annotazione. Il livello inferiore è suddiviso a sua volta in tre blocchi diversi che non dialogano tra loro bensì rispondono solamente al livello superiore. Questo permette di rendere indipendenti fra loro i moduli e di poter operare sul codice di ognuno senza dover modificare gli altri. L'ultimo modulo è quello che permette il benchmark dei risultati. Esso possiede funzionalità a sè stanti, di conseguenza non è collocabile a nessun livello architetturale. Il sistema opera con una serie di step successivi:

- Data Acquisition
- Query all'annotatore
- Discriminazione entità relative a luoghi
- Scelta del luogo migliore
- Data Saving
- Benchmarking analysis

[1]Per approfondire sulla tecnologia CGI: https://it.wikipedia.org/wiki/Common_Gateway_Interface

[2]Non solo è una forma di RPC bensì può essere costruita una vera e propria interfaccia RPC a partire da uno script CGI: http://www.ibm.com/developerworks/library/x-tipxmlrpc/

Nelle sezioni successive verranno descritti in linea di massima gli annotatori semantici utilizzati e inoltre verrà analizzato il lavoro effettuato step by step.

3.2 Annotatori semantici

3.2.1 Tagme

La prima versione di Tagme è stata rilasciata nel 2010. Tagme è stato pensato per lavorare con testi brevi (composti da qualche decina di termini) in modo molto veloce. La risposta fornita dallo strumento include per ogni annotazione un parametro chiamato rho. Questo parametro indica la "bontà" dell'annotazione in relazione al significato del testo nella sua interezza. Può essere usato quindi per scartare annotazioni sotto una certa soglia ed è l'equivalente della confidenza. Il parametro rho è sempre compreso nell'intervallo [0,1]. Viene consigliato di utilizzare 0.1 come soglia di attendibilità, anche se è opportuno eseguire alcuni test per poter trovare il valore migliore. Utilizzando dei parametri opzionali la risposta può includere alcune informazioni aggiuntive come la categoria della pagina associata all'annotazione (da DBpedia) oppure dei comportamenti specifici come nel caso di tweet. La richiesta deve essere inviata ad un server centralizzato con API di tipo REST. La risposta è fornita tramite un oggetto JSON.

3.2.2 DBpedia Spotlight

DBpedia Spotlight è uno strumento di annotazione testi che fornisce collegamenti a risorse DBpedia. Lo strumento fornisce quindi un meccanismo per annotare testi con informazioni non strutturate tramite DBpedia. Vengono messe a disposizione tre funzioni di base: Spotting semplice, Disambiguazione e Candidates [6]. È possibile accedere a questi servizi con API Scala/Java oppure tramite un approccio REST. Il contenuto della risposta proveniente dal Web Server può essere negoziato tramite la modifica di un parametro. Ad oggi sono supportati i seguenti tipi di risposta:

- text/html
- application/xhtml+xml
- text/xml
- application/json

3.2.3 Dexter

Dexter è un framework open source che fornisce tutti gli strumenti necessari per l'annotazione di testi. Dexter offre delle API REST con la possibilità di inviare richieste ad un server centralizzato oppure ad un server locale. Dexter è un servizio molto flessibile e altamente modulare; la risposta è fornita tramite un oggetto JSON. Lo strumento può essere regolato tramite un parametro, la confidenza minima (min-conf). Inoltre sia lo spotter (lo strumento che si occupa di dividere in spot) sia il disambiguatore possono essere modificati/sostituiti. Il servizio è gratuito e open source.

3.2.4 Confronto fra annotatori semantici

Features	**TagMe**	**Dexter**	**Spotlight**
Pricing	Free con chiave	Free e Open Source	Free e Open Source
Server	Centralizzato	Locale e Centralizzato	Locale e Centralizzato
Lingue supportate	IT, EN, DE	EN	Principali UE
Tuning	Rho (confidenza)	min-conf (confidenza)	min-conf (confidenza)
Spotter	Proprietario	Open e sostituibile	Open e sostituibile
Disambiguatore	Proprietario	Tagme ma sostituibile	Open e sostituibile

Tabella 3.1: *Confronto fra annotatori*

3.3 Data Acquisition

Il sistema non è provvisto di un vero e proprio meccanismo di *data acquisition*, al contrario accede ai dati dei tweets tramite un database relazionale MySQL. Lo strumento quindi richiede una fase preliminare in cui i tweets da analizzare vengono raccolti e messi a disposizione dello strumento. Naturalmente, a patto di fornire i dati all'interno di un database relazionale è possibile far lavorare lo strumento anche con uno stream di dati.

3.4 Query all'annotatore

In questa fase il testo dei tweets da analizzare viene passato ad uno dei tre moduli a livello più basso, in particolare a quello relativo all'annotatore selezionato. Il modulo si occupa di

interrogare lo strumento di annotazione e raccogliere in un oggetto JSON le risposte provenienti dall'annotatore. Tutti e tre gli annotatori vengono eseguiti su un server remoto ed espongono un'interfaccia di tipo RESTful. La funzione principale dei tre sottomoduli quindi è di uniformare il meccanismo di richiesta all'annotatore e restituire un oggetto che non varia a seconda dell'annotatore scelto.

Figura 3.2: *Metodi esposti da ogni modulo per l'interrogazione di un annotatore e oggetto di ritorno*

Nella figura infatti possiamo vedere come i metodi dei tre moduli richiedono il passaggio degli stessi argomenti. La risposta di ogni metodo è un oggetto JSON contenente tutte le entità annotate e non solo relative a luoghi geografici. La sola funzione del modulo infatti è quella di richiedere l'annotazione del testo. Successivamente il risultato dell'annotazione verrà elaborato per discriminare tutte le entità che non si riferiscono ad un luogo geografico. Inoltre per ogni entità taggata nel tweet viene restituito il link al database di conoscenza DBpedia e il parametro di confidenza dell'associazione. Mantenere per ogni associazione anche la confidenza della stessa potrebbe sembrare superfluo dato che il filtraggio viene effettuato già dall'annotatore, ma potrebbe essere utile comunque in fase di analisi dei risultati.

3.5 Discriminazione entità relative a luoghi

Tutte le entità relative ad un tweet vengono analizzate dal modulo che si occupa della discriminazione di categoria tramite interrogazione al database di conoscenza. Il modulo si fa carico di un'ulteriore funzione oltre a quella della discriminazione ovvero recuperare le coordinate geografiche relative alle entità che considera luoghi geografici. In questo modo fornisce al livello intermedio un vettore di coordinate geografiche relative a tutte le entità di tipo *place* fra quelle ricevute in input. Il limite di un approccio con query SPARQL ad un database di conoscenza emerge quando l'entità da discriminare non è un place ma un Point of Interest (POI) oppure un'entità localizzabile ma che non fa parte di queste categorie. In questi casi sarebbe necessario un altro tipo di approccio che viene rimandato ad estensioni future. L'implementazione di questo modulo quindi viene volutamente mantenuta molto semplice in modo da poter essere facilmente modificata da chi effettivamente utilizza lo strumento stesso. Tutte le richieste implementate dal modulo sono effettuate all'endpoint del database di conoscenza. Una nota sulla tecnologia SPARQL è consultabile nell'appendice di questo documento.

3.6 Scelta del luogo migliore

Per ogni tweet possono essere annotati più luoghi geografici. Questo fatto però potrebbe costituire un problema se il tweet deve essere mostrato su una mappa. Inoltre spesso per valutare tra di loro gli annotatori è necessario stabilire uno standard di confronto. Sorge quindi la necessità di scegliere un solo place fra quelli annotati all'interno del testo. Per i motivi sopraelencati, è presente un modulo che si occupa della scelta del luogo migliore fra quelli che vengono a lui forniti in input. I criteri per la scelta del *best place* possono essere vari e dipendenti dal fenomeno preso in analisi. In caso di sisma infatti può essere utilizzato come elemento di calcolo l'epicentro del terremoto che però perde di senso nel caso d'analisi di uno tsunami in quanto l'epicentro del fenomeno può avvenire a migliaia di kilometri di distanza dalla zona effettivamente colpita. Il modulo quindi è scritto in modo che l'implementazione fornita possa essere facilmente integrata sfruttando altri fattori. Senza alcuna integrazione, viene scelto come luogo migliore l'entità che, in termini di distanza geografica, possiede distanza minore dall'epicentro/dagli epicentri del fenomeno. È necessario porre particolare attenzione a come viene calcolata la distanza geografica stessa. La forma sferica del globo infatti non permette il calcolo della distanza fra due punti come verrebbe fatto in un piano. Viene utilizzata quindi la formula dell'emisenoverso [11] che consente di calcolare la distanza fra due punti su una superficie sferica.

3.7 Data Saving

Il task svolto in questo step richiede un leggero focus riguardo a come i dati verranno adoperati in futuro. Nel caso dell'analisi presa in considerazione in questo documento il modulo semplicemente salva la coordinata geografica (latitudine e longitudine decimale), il nome del luogo taggato dall'annotatore e l'id della pagina Wikipedia relativa al luogo stesso. I dati sono ridondati (id e nome del luogo) per poter accedere nuovamente in maniera efficiente alla risorsa sul database di conoscenza.

3.8 Benchmarking analysis

Lo step finale ovvero quello della valutazione dei risultati è naturalmente opzionale dato che non sempre (in tempi utili!) è possibile conoscere la verità relativa ad un dataset di tweets. Questo modulo quindi opera esternamente al sistema e volutamente non è collocato in nessun livello architetturale. Il task svolto da questo modulo opera in due sottotask successivi da svolgere in maniera consecutiva:

- Standardizzazione degli input
- Confronto

3.8.1 Standardizzazione degli input

Il primo sottotask, la standardizzazione degli input, è necessaria in quanto lo strumento, tramite il modulo di "Scelta del luogo migliore", seleziona solamente un place per tweet. Modificando il sistema di scelta potrebbero essere scelti più place per singolo tweet. Il benchmark quindi risulterebbe falsato se non si tenesse conto di questo fondamentale particolare. È necessario uniformarsi all'output del sistema, di conseguenza per poter completare questo job il modulo di benchmark potrebbe rivolgersi al modulo di "Scelta del luogo migliore" prendendo in prestito alcune delle sue funzionalità.

3.8.2 Confronto

Il secondo sottotask, il confronto, permette di misurare numericamente sia il numero di tweets taggati sia la corrispondenza di place, fornendo quindi due valori complementari fra loro. Evidentemente un task di benchmarking può coinvolgere una serie di parametri per poter rendere la valutazione del sistema veritiera. A questo proposito, nel capitolo successivo sarà possibile approfondire questo aspetto.

4. Esperimenti

4.1 Premessa

La valutazione delle risposte del sistema è stata effettuata tramite due diversi casi di studio. Nel primo caso il dataset sottoposto ad annotazione contiene dei tweets relativi al Terremoto dell'Emilia (2012), all'alluvione della Sardegna (2013) e agli attentati di Parigi (2015). Nel secondo caso invece è stato preso in analisi un evento che ha tristemente segnato l'Italia nel mese precedente ovvero il Terremoto del Centro Italia (2016). Evidentemente i tweets, in entrambi i casi di studio, sono solamente una piccola parte rispetto alla totalità dei dati che il social media mette a disposizione. In particolare per il terremoto del Centro Italia sono stati presi in considerazione tutti i tweets relativi alla prima ora post-evento.

La differenza sostanziale, tale per cui i casi di studio sono stati divisi in due categorie (e non in quattro come farebbe pensare immediatamente il numero degli eventi presi in considerazione!), è la modalità con cui è stata calcolata la verità. Nel primo caso infatti è stata necessaria una fase preliminare in cui sono stati analizzati i tweets annotando manualmente i vari luoghi presenti in ogni messaggio. Nel secondo caso del fenomeno analizzato non si possiedono dati certi per cui la valutazione è qualitativa e si basa su una parte di dati ufficiali della protezione civile riguardo alle zone colpite dal sisma. In questo modo abbiamo cercato quindi di fornire una valutazione sia qualitativa, cercando di scoprire quanto lo strumento funziona bene in una situazione reale e tempestiva, sia quantitativa, andando a verificare quanto la capacità di leggere, comprendere e annotare dello strumento sia vicina a quella posseduta da un essere umano. Per il secondo caso di studio sono state generate delle mappe di crisi in modo da poter analizzare i risultati in maniera visuale. In appendice è presente un paragrafo relativo alle mappe di crisi e a come possono essere utilizzate nella valutazione di task di geoparsing.

Dalla tabella si nota subito una differenza di circa un ordine di grandezza per quanto riguarda il numero di tweets dei due diversi casi di studio. Questa differenza è giustificata dal fatto che nel primo caso di studio i tweets posseggono sicuramente almeno un place per tweet dato che sono stati annotati a mano. Nel secondo caso invece sono presenti molti tweets che

	Primo caso	**Secondo caso**
Emilia 2012	534	-
Sardegna 2013	274	-
Parigi 2015	399	-
Centro Italia 2016	-	10226
Totale	1207	10226

Tabella 4.1: *Numero di tweets analizzati per evento e caso di studio*

non contengono alcuna informazione.

4.2 Primo caso di studio

La valutazione delle risposte dello strumento è stata testata in questo caso di studio su tweets dei seguenti fenomeni:

- Alluvione in Sardegna (anno 2013)
- Terremoto in Emilia (anno 2012)
- Attacchi terroristici di Parigi (anno 2015)

Il dataset quindi è molto vario e si adatta bene ad uno studio valutativo. I tweets relativi agli attentati di Parigi inoltre sono in lingua inglese. Il dataset è stato analizzato e geotaggato manualmente, si possiede quindi una *ground-truth*[1]. In questo caso sono stati presi in considerazione tutti i place trovati all'interno del tweet e non solo quello migliore. L'analisi dei risultati può essere espressa in termini statistici ed, in particolare, tramite una *Confusion matrix*[2]. La matrice è composta da quattro celle, *true positive*, *true negative*, *false positive*, *false negative*. Il valore nella cella true positive rappresenta il numero dei place taggati manualmente dall'annotatore e, parallelamente classificati con successo anche dallo strumento; nella cella true negative è presente il numero delle entità che non riguardano place e che correttamente l'annotatore non ha identificato nella ricerca; il valore false positive rappresenta un falso allarme dello strumento in quanto ha riconosciuto come place un'entità che non è

[1]Una verità di fondo di cui siamo statisticamente certi: https://en.wikipedia.org/wiki/Ground_truth

[2]Utilizzata nell'ambito dell'intelligenza artificale restituisce una rappresentazione dell'accuratezza di una scelta di classificazione: https://en.wikipedia.org/wiki/Confusion_matrix

tale; per ultimo il valore false negative rappresenta un place riconosciuto come tale dalla verità ma non associato dall'annotatore. Tenendo conto di questi quattro valori quindi vengono calcolate alcune metriche che rappresentano la bontà dell'annotatore nel riconoscimento dei luoghi.

		Annotatore	
		Place	Non Place
Verità	Place	True Positive	False Negative
	Non Place	False Positive	True Negative

Tabella 4.2: *Descrizione valori della confusion matrix*

		Annotatore	
		Place	Non Place
Verità	Place	1225	355
	Non Place	191	13936

(a) *Tagme*

		Annotatore	
		Place	Non Place
Verità	Place	1221	359
	Non Place	212	13915

(b) *Spotlight*

		Annotatore	
		Place	Non Place
Verità	Place	1333	247
	Non Place	607	13520

(c) *Dexter*

Tabella 4.3: *Confusion matrix relative ai tre annotatori (confidenza 0.1)*

	Tagme	**Spotlight**	**Dexter**
Precision	0.87	0.85	0.69
Recall	0.78	0.77	0.84
F1 score	0.82	0.81	0.76

Tabella 4.4: *Valori relativi al caso migliore (confidenza 0.1) dei tre annotatori*

Le tabelle visibili sopra rappresentano la risposta dello strumento con una confidenza di 0.1. I valori di *precision* e *recall* ci indicano che il comportamento di Tagme e Spotlight è equivalente. Dexter invece tenta di associare molto più spesso degli altri due, per cui è poco

preciso ma contemporaneamente riconosce molti più place. Queste analisi sono fatte sui dati ottenuti mantenendo 0.1 di confidenza. Questo però potrebbe non essere sufficiente a definire la qualità della risposta stessa. L'analisi quindi è stata approfondita tramite la modellazione di una curva ROC (Receiver operating characteristic)[3]. Lungo l'asse delle ordinate viene rappresentato il *True Positive Rate*[4] ovvero la probabilità che un place venga riconosciuto con successo dall'annotatore.

$$TPR = TP/(TP + FN)$$

Evidentemente se il TPR fosse 1, significherebbe che l'annotatore riconosce con successo ogni place all'interno di un tweet. Lungo l'asse delle ascisse viene rappresentato il *False Positive Rate*[5] ovvero la probabilità che l'annotatore identifichi come place un'entità che non lo è.

$$FPR = 1 - TN/(FP + TN)$$

Evidentemente se il FPR fosse 0, significherebbe che l'annotatore non riconosce mai come place un'entità che non lo è. La curva ROC è l'insieme delle coppie TPR, FPR al variare di un parametro classificatore che in questo caso è la confidenza restituita dall'annotatore. In questo modo possiamo rappresentare le informazioni della matrice al variare della soglia di confidenza.

Dalle curve ROC possiamo vedere che entrambi gli annotatori funzionano meglio con una confidenza piuttosto bassa, e per questo motivo nelle matrici viste sopra i dati sono relativi ad una confidenza pari a 0.1. L'analisi del dato è piuttosto semplice, abbassando la confidenza vengono identificate numerose entità in più andando ad incrementare sia i place trovati con successo (TP) sia i falsi allarmi (FP). Nello stesso modo vanno a decrementare le altre due grandezze della matrice (ovvero FN e TN). Per come sono calcolati i rapporti, il valore che indica la probabilità di trovare con successo un place aumenta mentre rimane pressoché invariato quello che indica il fallimento da parte dell'annotatore (il valore TN è maggiore di FP di circa un ordine di grandezza!). Inoltre la curva è piuttosto ripida e si avvicina molto al punto (0:1) ovvero la situazione di ottimalità. Queste curve ci forniscono alcune informazioni importanti, ovvero gli strumenti di annotazione si avvicinano alla perfezione e riescono ad estrarre da un tweet quasi ogni informazione che si riferisce ad un luogo geografico.

---

[3]Nella teoria delle decisioni le curve ROC sono schemi per caratterizzare dei classificatori binari, per approfondire: https://it.wikipedia.org/wiki/Receiver_operating_characteristic

[4]Il TPR, detto anche *Sensitivity*, è l'equivalente del Hit-rate o Recall

[5]Il FPR è l'equivalente del Fall-out

4.3 Secondo caso di studio

In questo caso di studio sono stati presi in considerazione i tweets relativi al terremoto del centro italia. Il fenomeno avvenuto è recente alla scrittura di questo documento per cui non si possiede una ground truth. Il confronto quindi è basato sui dati forniti fino ad adesso dalle autorità anche se gli stessi non sono definitivi. Le mappe coropletiche in questo caso ci possono guidare alla valutazione delle risposte dello strumento e sono disponibili al seguente url: http://wafi.iit.cnr.it/crismap/crismap/mapping/. Le mappe sono state calcolate per diversi livelli di zoom. Partendo dalla mappa relativa alle regioni si nota che la regione più colpita (e quindi di colore rosso) è l'Umbria. Aumentando il livello di zoom possiamo osservare che le province più intensamente colorate (Spoleto, Terni, Rieti, ..) sono concentrate nei pressi dell'epicentro del fenomeno. L'ultimo livello di zoom disponibile ci fornisce un dato fondamentale. I comuni in cui è stato riscontrato l'epicentro del terremoto infatti sono stati colorati di rosso intenso. Questo significa che il fenomeno è stato localizzato correttamente.[6]

[6]La valutazione in questo caso si è basata su dati non ancora definitivi provenienti dal blog ufficiale dell'Istituto Nazionale di Geofisica e Vulcanologia: https://ingvterremoti.wordpress.com/

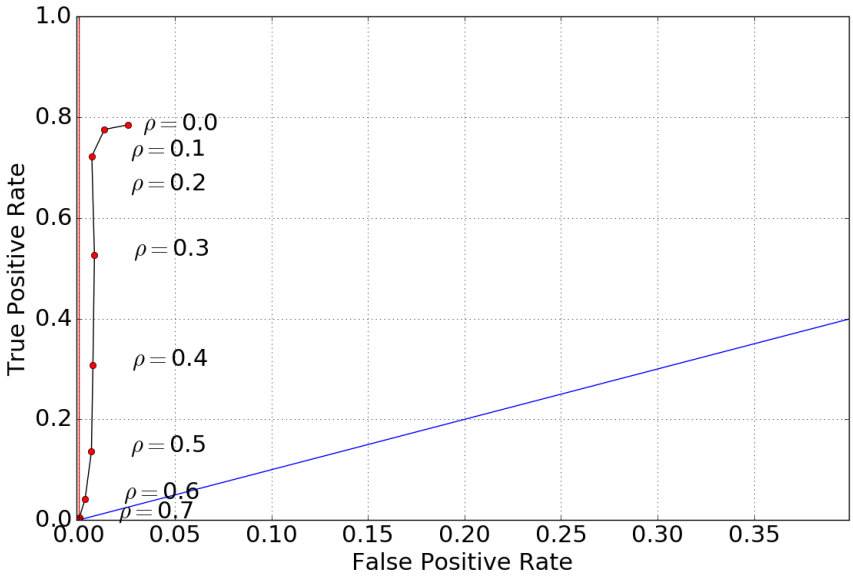


Figura 4.1: *Curva ROC relativa all'annotatore Tagme, la bisettrice blu rappresenta il comportamento di un classificatore randomico, l'asse rosso il comportamento di un classificatore perfetto.*

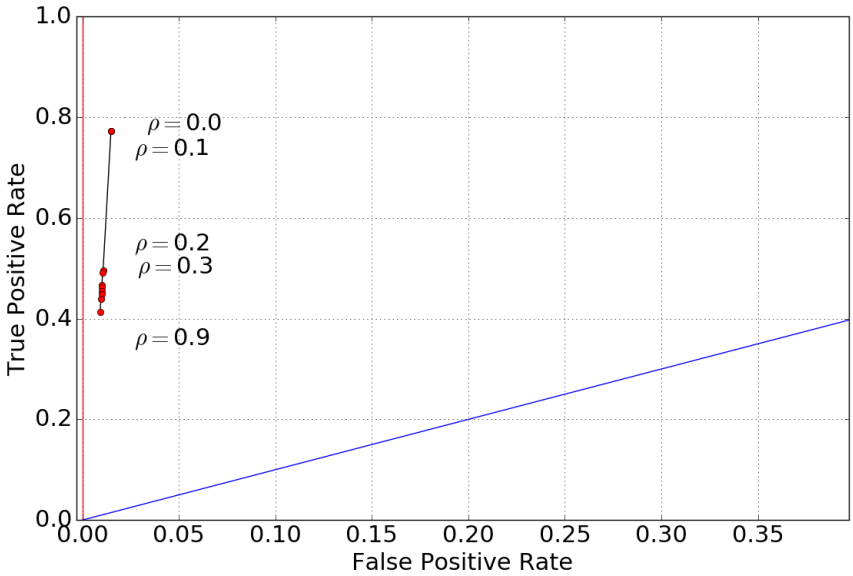


Figura 4.2: *Curva ROC relativa all'annotatore Spotlight, la bisettrice blu rappresenta il comportamento di un classificatore randomico, l'asse rosso il comportamento di un classificatore perfetto.*

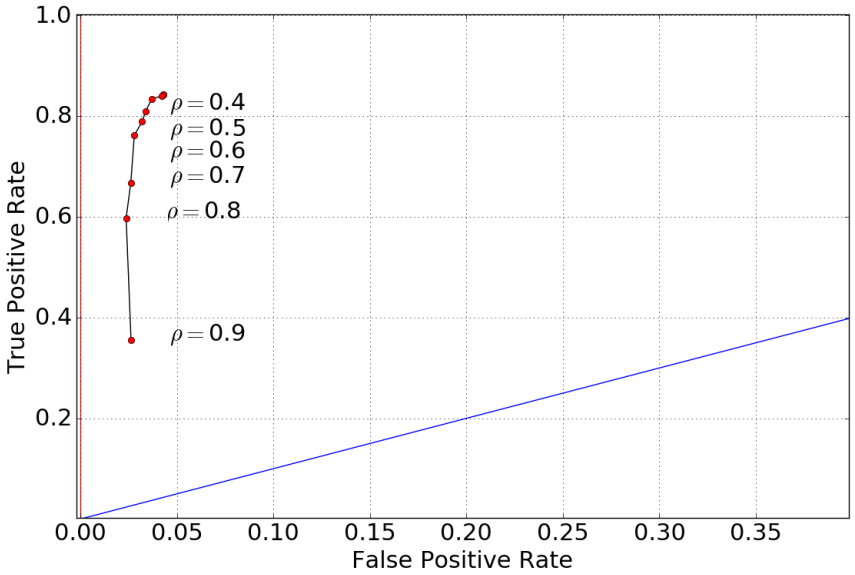


Figura 4.3: *Curva ROC relativa all'annotatore Dexter, la bisettrice blu rappresenta il comportamento di un classificatore randomico, l'asse rosso il comportamento di un classificatore perfetto.*

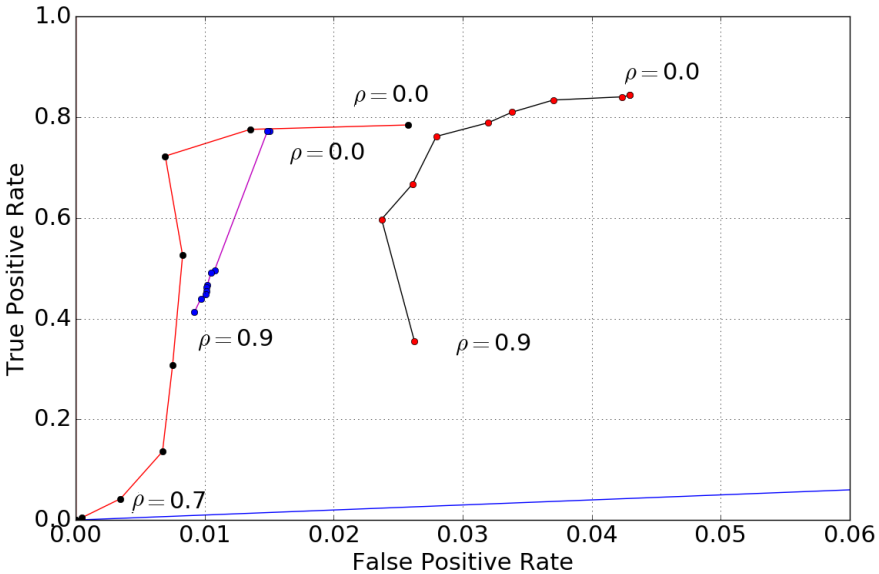


Figura 4.4: *Confronto delle curve ROC sullo stesso piano. In rosso la curva ROC relativa all'annotatore Tagme, in nero la curva ROC relativa all'annotatore Dexter, in magenta la curva ROC relativa all'annotatore Spotlight. La bisettrice blu rappresenta il comportamento di un classificatore randomico.*

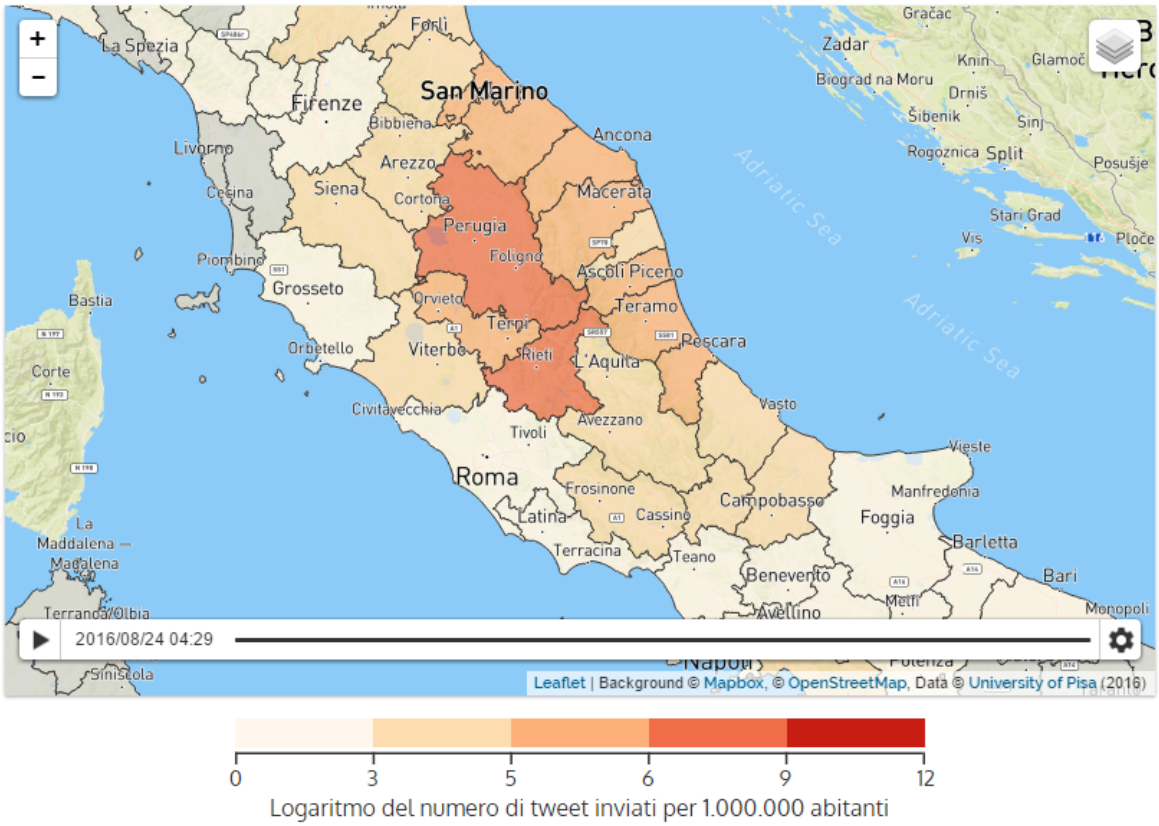
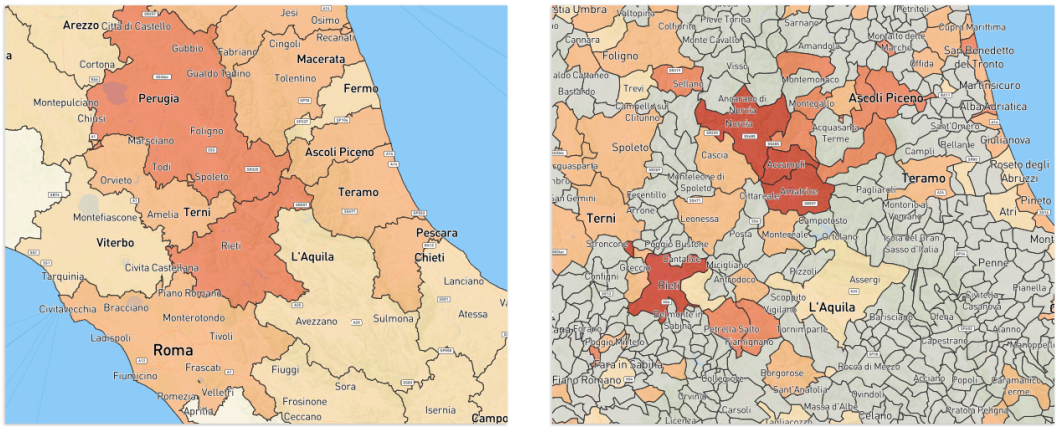


Figura 4.5: *Mappa coropletica relativa al terremodo del Centro Italia (2016)*



(a) **(b)**

Figura 4.6: *Stessa mappa per diversi livelli di zoom. Si può notare il diverso livello di dettaglio, nella figura (a) sono rappresentati i dati aggregati per provincia, nella figura (b) i dati relativi ai singoli comuni nei pressi dell'epicentro.*

5. Conclusioni

Il sistema sviluppato e i risultati ottenuti sui due casi di studio analizzati nel capitolo 4 ci consentono di affermare che l'approccio seguito è valido e già realmente utilizzabile.

Tramite il primo caso di studio è stato possibile valutare la precisione con cui gli strumenti di annotazione riescono a leggere un testo in linguaggio naturale e collegarlo alle rispettive entità. Buona parte degli errori causati dagli annotatori sono legati alla scarsa lunghezza del testo da analizzare che non consente una piena comprensione da parte dello strumento. Altri invece sono causati da abbreviazioni introdotte dagli utenti oppure frazioni di comuni che non posseggono la relativa voce sul database di conoscenza. Il cospicuo numero di tweets però permette di localizzare con una buona precisione i luoghi colpiti dal fenomeno. Nell'ottica di una futura implementazione potranno essere presi in considerazione anche i dati personali degli utenti che hanno scritto il tweet (ricavando così luogo di origine, luogo di residenza e altri dati chiave per l'analisi). Inoltre un percorso da seguire sarà quello di aumentare il numero di annotatori, facendoli lavorare non come strumenti a sé stanti bensì a stretto contatto con un continuo scambio di informazioni.

Mediante il secondo caso di studio è stato mostrato come già nella prima ora successiva al terremoto avessimo a disposizione una mappa generale che rappresenta fedelmente i dati provvisori registrati dall'Istituto Nazionale di Geofisica e Vulcanologia.

Il sistema inoltre è stato progettato per poter evolversi in direzione di un utilizzo con stream di dati real-time. L'analisi di tweets tramite annotatori ospitati su server remoti è molto dispendiosa in termini di tempo e spesso il numero di tweets raccolti supera quello che la macchina riesce ad analizzare; la maggior parte del tempo però è causato dalla richiesta/-risposta al server e non dall'elaborazione perciò un possibile sviluppo riguarda la possibilità di ospitare localmente gli annotatori semantici che lo prevedono.

Appendice A. Linguaggio SPARQL

SPARQL[1] è un query language che permette di ottenere informazioni attraverso numerose sorgenti di dati ed è uno degli elementi che caratterizza oggigiorno la maggioranza delle tecnologie legate al paradigma del *web semantico*. Con web semantico si intende un ambiente in cui i documenti pubblicati sono associati a informazioni e metadati, organizzati secondo il loro significato semantico in un formato adatto all'interrogazione e alla loro corretta interpretazione.[2] Le voci DBpedia infatti sono descritte tramite RDF (Resource Description Framework) che descrive i concetti e le relazioni su di esse attraverso l'introduzione di triple (soggetto-predicato-oggetto). Il linguaggio d'interrogazione SPARQL consente la costruzione di query basate su triple patterns, congiunzioni logiche, disgiunzioni logiche, e pattern opzionali. La query utilizzata per ottenere informazioni relative alle coordinate geografiche è visibile nello snippet di codice sottostante

```
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbont: <http://dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?label ?id ?lat ?lon
WHERE { <entityURI>
            rdfs:label ?label;
            dbont:wikiPageID ?id;
            geo:lat ?lat;
            geo:long ?lon.
}
```

Nel corpo del costrutto WHERE è visibile la struttura a tripla che caratterizza il linguaggio.

[1] È possibile trovare le specifiche all'url https://www.w3.org/TR/rdf-sparql-query/

[2] Il concetto può essere approfondito: https://it.wikipedia.org/wiki/Web_semantico

Appendice B. Mappe di Crisi

Le mappe di crisi permettono di visualizzare tramite colori e/o simboli grafici di immediata comprensione i danni avvenuti a cose e persone in una determinata area. Spesso mappe di questo tipo vengono prodotte partendo da dati provenienti dalla protezione civile o da altri fonti ufficiali. Frequentemente però le mappe fornite dalla protezione civile sono generate a partire da dati provenienti da stazioni sismiche automatiche, e per tali motivi non sono complete [2]. Lo stesso difetto evidentemente può emergere anche nel caso in cui le mappe vengano generate a partire da dati di social media, quindi ha senso utilizzare diverse fonti di informazioni.

Ci sono numerosi tipi di mappe di crisi, quelle visibili in questo documento sono di tipo coropletico. Inoltre questo tipo di mappe sono le più comuni fra le mappe tematiche [12] e di conseguenza le più documentate in letteratura [13].

Nelle mappe coropletiche ad ogni area viene associata un colore oppure una sua diversa sfumatura in proporzione al valore che deve essere mostrato. Tipicamente la mappa viene suddivisa in zone con confini più o meno definiti; spesso si utilizzano quelli territoriali di regioni, province, stati. Le mappe di questo tipo quindi rappresentano informazioni aggregate creando di fatto un compromesso tra dettaglio e semplicità.

In generale le mappe di crisi sono fondamentali come visualizzazione di un task di geoparsing. Una volta raccolte tutte le informazioni relative ai tweets letti dal sistema, le mappe di crisi permettono di analizzarle per trarre alcune immediate conclusioni su quali siano le aree effettivamente colpite e la loro condizione. In questo modo è possibile organizzare eventuali soccorsi in modo da ottimizzare il loro utilizzo e intervenire in modo più tempestivo. Unire le informazioni a disposizione degli enti con quelle ottenute in questo modo permette di avere una chiara conoscenza di quello che sta accadendo e della situazione del territorio. Questo permette di evitare inutili dispendi di energie e spesso di salvaguardare la vita stessa dei soccorritori, evitando loro inutili pericoli. La presentazione delle informazioni quindi è un task fondamentale e deve essere il più possibile di immediata comprensione.

Bibliografia

[1] A. Rosi, S. Dobson, M. Mamei, G. Stevenson, J. Ye, and F. Zambonelli, "Social sensors and pervasive services: Approaches and perspectives." in *Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, 2011, pp. 525–530.

[2] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi, "EARS (Earthquake Alert and Report System): A real time decision support system for earthquake crisis management," in *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2014, pp. 1749–1758.

[3] M. Avvenuti, S. Cresci, F. Del Vigna, and M. Tesconi, "Impromptu crisis mapping to prioritize emergency response," *Computer*, vol. 49, no. 5, 2016.

[4] P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1625–1628. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871689

[5] S. Trani, D. Ceccarelli, C. Lucchese, S. Orlando, and R. Perego, "Dexter 2.0 - an open source tool for semantically enriching data," in *Proceedings of the 13th International Semantic Web Conference*, ser. CIKM '10, 2014.

[6] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[7] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[8] R. Usbeck, M. Röder, A.-C. N. Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli,

F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann, "Gerbil – general entity annotation benchmark framework," in *Proceedings of the 24th WWW conference*, 2015.

[9] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," in *IEEE Intelligent Systems (Volume: 29, Issue: 2, Mar.-Apr. 2014)*, ser. CIKM '10. IEEE, 2013, pp. 9 – 17.

[10] E. Loper and S. Bird, "NLTK: the natural language toolkit," *CoRR*, vol. cs.CL/0205028, 2002. [Online]. Available: http://arxiv.org/abs/cs.CL/0205028

[11] C. Veness, "Calculate distance and bearing between two latitude/longitude points using haversine formula in javascript," *Movable Type Scripts*, 2011.

[12] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard, "Thematic cartography and geovisualization," 2009.

[13] K. Goldsberry and S. Battersby, "Issues of change detection in animated choropleth maps," vol. 44, no. 3. International Cartographic Association/Association Cartographique internationale, 2009, pp. 201–215.