UNIVERSITÀ DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI INFORMATICA
CORSO DI LAUREA TRIENNALE IN INFORMATICA

TESI DI LAUREA IN METODI PER IL
RITROVAMENTO DELL'INFORMAZIONE

SISTEMI DI SUPPORTO ALLE DECISIONI BASATI SULL'ELABORAZIONE DI DATI TESTUALI E METADATI PER LA GESTIONE DELLE POLITICHE DI INNOVAZIONE

RELATORE
Chiar.mo Prof. Pasquale LOPS
Dott. Pierpaolo BASILE

LAUREANDO
Flavio PETRUZZELLIS

ANNO ACCADEMICO 2018/2019

Indice

Introduzione

I sistemi di supporto alle decisioni sono sistemi informatici che hanno lo scopo di aumentare l'efficacia e l'efficienza dei processi decisionali in qualunque tipo di organizzazione. Il valore aggiunto apportato da questi sistemi ai processi decisionali consiste nell'offrire agli stakeholder la possibilità di tenere in considerazione informazioni e conoscenze relative al dominio di interesse che non sarebbero disponibili senza di essi, per poter compiere decisioni o organizzare processi in contesti parzialmente o completamente variabili e di difficile prevedibilità. I processi di ritrovamento e di elaborazione dell'informazione sono svolti in modo automatico a causa della grande quantità di informazioni e conoscenze prodotte nelle organizzazioni di medie e grandi dimensioni, che è, peraltro, conservata sempre più frequentemente solo in formato digitale.

La maggior parte dei sistemi di supporto alle decisioni impiegano risorse informatiche per elaborare informazioni strutturate, ossia dati che descrivono il dominio di interesse che solitamente conservati in grandi basi di dati o *data warehouses*. Queste analisi sono svolte tramite tecniche di *data mining* che hanno il fine di estrarre conoscenza precedentemente ignota a partire dai dati, applicando metodi algoritmici o statistici.

Una tipologia alternativa di sistemi di supporto alle decisioni è quella basata sull'estrazione di informazioni e conoscenza a partire dai dati non strutturati, ossia dai comuni testi scritti in linguaggio naturale. I sistemi più diffusi in questa categoria sono indubbiamente i motori di ricerca, che hanno l'obiettivo di reperire e mostrare all'utente i documenti di una collezione che sono rilevanti rispetto ad un particolare bisogno informativo. Il criterio con il quale questo genere di sistemi valuta la corrispondenza tra un documento e un bisogno informativo si basa solitamente sulla

valutazione dell'occorrenza nel documento di specifiche parole con cui esso è espresso. È anche possibile, però, applicare ai documenti tecniche di elaborazione che tengono in considerazione il significato dei termini. Attraverso queste tecniche è possibile sia creare motori di ricerca più sofisticati, sia progettare e sviluppare servizi basati sulla rappresentazione semantica del testo. Questo genere di servizi dà all'utente la possibilità di esaminare il contenuto dei documenti e fare confronti su di essi sfruttando il processo di analisi automatica. Ad esempio, è possibile individuare i documenti in una collezione che trattano di un tema, individuare i concetti simili tra loro sulla base di come vengono usati all'interno di una collezione, o, ancora, trovare i documenti di una collezione che parlano di temi simili.

Il presente lavoro di tesi è incentrato sullo studio del funzionamento e delle possibilità di ampliamento di un sistema di supporto alle decisioni basato su dati non strutturati, sviluppato nell'ambito del progetto europeo TALIA, sfruttando i servizi offerti da un sistema di gestione documentale e analisi semantica del testo chiamato Semantic Framework. In particolare, è obiettivo della tesi indagare le possibilità di miglioramento dei servizi offerti dal sistema di supporto alle decisioni mediante l'integrazione in essi di informazioni strutturate riguardanti il dominio di interesse.

Il lavoro di tesi è così strutturato: il primo capitolo ha l'obiettivo di illustrare i principi teorici e le tecniche sulla base delle quali è stato sviluppato il Semantic Framework. Nel secondo capitolo viene presentato il progetto TALIA e il suo dominio, ambito di applicazione del sistema di supporto alle decisioni che è stato oggetto dello studio. Vi sono inoltre introdotti gli specifici servizi offerti dal sistema di supporto alle decisioni. Il terzo capitolo illustra le scelte progettuali compiute per integrare nel Semantic Framework un modulo di gestione dei dati strutturati relativi al dominio

cui appartengono i documenti gestiti e analizzati nella piattaforma. In conformità con l'architettura del sistema, i servizi aggiuntivi sono stati sviluppati secondo il protocollo REST, del quale sono illustrati i principi di funzionamento. Il quarto capitolo illustra i risultati dell'integrazione dei dati strutturati nei servizi del sistema di supporto alle decisioni, supportati da evidenze sperimentali. Il sesto capitolo presenta, infine, conclusioni e sviluppi futuri del progetto.

1. Sistemi basati su dati testuali

La maggior parte delle informazioni attualmente disponibili in formato digitale, siano esse di pubblico dominio o appartenenti a un'organizzazione, sono organizzate in forma non strutturata, ossia sottoforma di testi o frammenti di testo.

Dal punto di vista della rappresentazione della conoscenza, il testo è un formato radicalmento diverso rispetto a quello strutturato con cui sono memorizzate le informazioni in una base di dati, o rispetto alle regole logiche con le quali è codificata la conoscenza su un dominio in una base di conoscenza. Il testo è, infatti, una forma di rappresentazione della conoscenza prodotta da e per le persone, perciò esso va processato in modo opportuno affinché il suo contenuto informativo possa essere acquisito, processato e reso disponibile automaticamente, e non soltanto tramite la lettura. Esistono due livelli di rappresentazione digitale del testo: il più superficiale è basato solo sulla *forma* del testo e non sul suo *significato* (ossia, ad esempio, le parole "cappello" e "berretto", pur avendo un significato molto simile, sono considerate diverse perché hanno una forma diversa); diversamente da questo, il livello di rappresentazione dei testi più profondo e completo tiene conto del significato delle parole.

Le possibilità aperte dal poter operare automaticamente sui testi sono significative: diventa possibile gestire grandi collezioni di testi, contenenti anche decine di milioni di parole. Tramite i motori di ricerca è possible reperire documenti utili al loro interno cercando tramite parole chiave o a partire da "concetti", se la rappresentazione del testo ne preserva la semantica.

1.1 Elaborazione del linguaggio naturale

I passi compiuti per preparare il testo ad essere rappresentato in forme appropriate per le operazioni di ritrovamento dell'informazione fanno parte di un processo detto di "elaborazione del linguaggio naturale" o Pipeline di Natural Language Processing (NLP). Esso comprende un insieme di passi che hanno il fine di aumentare l'efficacia delle tecniche di rappresentazione e analisi del testo, tanto eliminando le ridondanze tipiche di questo tipo di rappresentazione della conoscenza, quanto mettendone in risalto la peculiare ricchezza informativa. Le operazioni eseguite in questo processo sono alla base sia della rappresentazione "superficiale" del testo, sia di quella che tiene conto del significato dei termini.

1.1.1 Sentence Detection

Il primo passo nel processo di elaborazione del testo è l'individuazione delle frasi. Esse vengono individuate all'interno di un testo considerando che sono solitamente separate dal punto fermo. Tuttavia, è necessario distinguere questo da altri usi del punto che possono essere presenti nei testi considerati (ad esempio, per abbreviare termini, per separare porzioni di un numero, eccetera).

1.1.2 Tokenization

Il secondo passo nella pipeline di NLP è la suddivisione delle frasi in parole o "token", eliminando i segni di punteggiatura e i caratteri speciali. Anche in questo caso, l'euristica base con la quale si identificano i token è che essi sono generalmente separati da spazi; bisogna tuttavia tener conto di token in forme particolari, come le date e l'ora, i nomi propri di persona, ecc., che sono composti da più parti che non vanno considerate separatamente.

1.1.3 Stop Words Removal

Le *stop words* sono parole molto comuni in una lingua o in un particolare ambito, al punto tale che il loro contenuto informativo è considerato molto basso o nullo; per questa ragione un testo si può rappresentare efficacemente senza comprenderle risparmiando spazio di rappresentazione e tempo di computazione. Per le lingue più comuni esistono liste di *stop words* che vengono usate per filtrare le parole rilevanti nei documenti.

1.1.4 POS Tagging

Il Part Of Speech Tagging è l'operazione che associa a ciascun token la sua categoria grammaticale all'interno della frase. Questa operazione non è banale, poiché molto spesso uno stesso termine può avere più di un significato, o perché alcuni elementi sintattici nella frase sono sottointesi nell'uso comune della lingua. La disambiguazione tra i diversi ruoli sintattici e l'individuazione di quello corretto avviene considerando sia le definizioni del termine considerato, che, soprattutto, le parole assieme alle quali occorre e il loro ruolo all'interno della frase.

L'operazione di POS Tagging è particolarmente utile per la costruzione di una rappresentazione della sintassi del testo e per l'individuazione della corretta semantica associata ai termini, nel caso in cui questi abbiano un diverso significato a seconda del loro ruolo sintattico all'interno della frase.

1.1.5 Lemmatization & Stemming

Nei documenti sono spesso presenti diverse forme di una parola che hanno dei significati simili (democrazia, democratico, democratizzazione). La lemmatizzazione è un'operazione di semplificazione dei token che trasforma ciascuno di essi nella sua forma grammaticale di base (ossia il suo

lemma) trasformando il suffisso proprio della specifica inflessione presente nel testo in quello della forma base dell'elemento grammaticale. Il processo di trasformazione non è banale, in quanto va identificata correttamente la forma base di ogni token; in particoalare, ciò è più difficile nel caso di termini che hanno la stessa forma ma funzioni grammaticali e significati differenti (es. un'àncora, egli àncora, ancòra qui). In questi casi gli algoritmi di lemmatizzazione possono ricondurre il termine al lemma corretto grazie al contesto in cui occorre e al risultato del processo di POS tagging.

Lo stemming è un'operazione di semplificazione dei token più radicale: essa tronca il suffisso di ogni token riducendo il termine alla sua radice. Uno degli algoritmi più usati per compiere questa operazione è l'algoritmo di Porter [1].

1.1.6 Keyphrase Detection

Questo passo di analisi del testo consiste nell'identificare i concetti più rilevanti all'interno di un testo, che ne descrivono dunque l'argomento e il contenuto informativo; essi possono essere espressi anche con più di una parola (ad es. "information retrieval"). Un approccio semplice, che impiega una tecnica di apprendimento non supervisionato, si basa sull'identificazione delle coppie, triple o n-uple di termini che co-occorrono con una frequenza significativa [2].

1.1.7 Chunking

Il Chunking è un'operazione di analisi sintattica del testo che può essere vista come un'operazione di tokenizzazione più elaborata, o come un parsing più semplice (è, infatti, anche detta "shallow parsing"). Essa consiste nell'individuare gli elementi grammaticali che sono composti da più termini, come i nomi propri di luoghi o di persone o le forme verbali

che comprendono verbi ausiliari o modali. Ciò è utile perché è più appropriato considerare queste forme grammaticali in modo unitario per avere contezza del loro significato; ad esempio, dire che in un testo si parla di Sud Africa è diverso dal dire che si parla di Sud e di Africa.

1.2 Motori di ricerca classici

Uno degli strumenti più comuni e più utili che si può costruire su una collezione digitale di documenti, processati attraverso le operazioni appena descritte, è il motore di ricerca. Esso è un sistema che, dato un insieme di parole – detto "query" – che esprime un bisogno informativo dell'utente, ha l'obiettivo di restituire ad esso i documenti della collezione che sono più rilevanti rispetto ai termini della query. Questo task cambia significativamente in relazione al tipo di rappresentazione del testo adottata. Nel caso in cui i documenti siano rappresentati tenendo conto della semantica dei termini al loro interno, infatti, i documenti più rilevanti per una query saranno quelli che contengono parole dal significato più simile a quelle date in input. Nel caso più semplice in cui i documenti siano rappresentati senza tener conto del significato dei termini, il calcolo della similarità tra documenti e query si basa sull'occorrenza delle parole della query nei documenti, ed eventualmente sulla frequenza con cui queste occorrono.

In questo paragrafo sono considerati i motori di ricerca "classici", ossia basati solo sulla forma dei termini e non sul loro significato. Esistono diversi approcci per creare un motore di ricerca classico; essi differiscono principalmente per il modo in cui sono rappresentati i documenti all'interno del sistema e per il modo in cui viene calcolata la similarità tra essi e le query dell'utente. Sono presentate di seguito le strutture dati e le

assunzioni sulla rappresentazione dei documenti comuni a diversi tipi di motori di ricerca; viene poi brevemente presentato il modello Booleano e il più diffuso Modello a Spazio Vettoriale.

1.2.1 Indice invertito

L'operazione più basilare ed intuitiva che un motore di ricerca classico deve compiere per risolvere una query è individuare quali sono i documenti della collezione che contengono le parole presenti in essa. L'insieme delle parole presenti in tutti i documenti di una collezione, associati alla loro frequenza assoluta, è detto "vocabolario" della collezione. Per compiere l'operazione di ritrovamento dei documenti in modo efficiente i motori di ricerca operano su una struttura dati chiamata "indice invertito". Il suo nome deriva dal confronto con il normale indice di un libro, che mette in relazione i capitoli o i paragrafi di un testo con le pagine in cui si trovano, e quindi con le parole contenute in essi. L'indice invertito è, invece, una struttura dati che associa ad ogni parola del vocabolario i documenti nei quali essa compare, consentendo così di individuare i sottoinsiemi di documenti in cui sono presenti tutte o molte delle parole in una query.

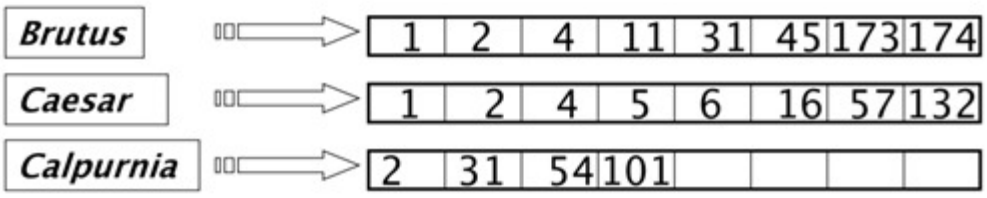

Figura 1.1 Un esempio di indice invertito, nel quale i numeri identificano i documenti della collezione

1.2.2 Modello Bag of Words

Il modello Bag of Words è un metodo di rappresentazione di un testo che tiene conto del numero di occorrenze di ogni parola nel testo ma non delle posizioni che ciascuna occupa in esso. Ciò implica che due o più testi che contengono le stesse parole ma in diverso ordine, e che hanno quindi complessivamente significati diversi, saranno rappresentati nello stesso modo. Questo modello si basa sull'assunzione forte che la posizione di una parola nelle frasi del testo in cui compare non sia rilevante per stabilirne il significato e, dunque, per interpretare il contenuto informativo del testo. Sebbene questa assunzione sia errata, e si punti, infatti, a sopperire alle sue mancanze tramite le operazioni di analisi sintattica del testo, come il Chunking e il POS Tagging, il motivo per cui si adotta è che la posizione delle parole in un testo non risulta empiricamente rilevante per il ritrovamento dei documenti attinenti alla richiesta espressa con una query.

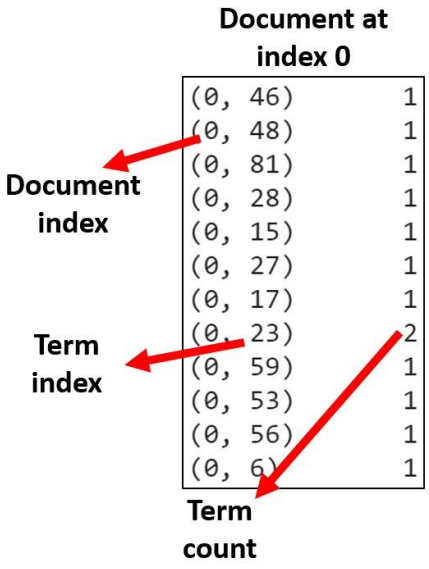

Figura 1.2 Rappresentazione secondo il modello Bag of Words di un documento identificato dall'indice 0

1.2.3 Il Modello Booleano

Adottando una rappresentazione conforme al modello Bag of Words, un documento sarà dunque rappresentato come un multiinsieme di parole, ossia un insieme nel quale gli elementi possono apparire più volte. Il modello booleano [3, 4] implementa nel modo più semplice possibile questa rappresentazione, considerando, cioè, la semplice presenza o assenza di un termine in un documento senza considerarne la frequenza d'uso. Un documento è rappresentato, quindi, come un vettore di variabili booleane, nel quale ogni elemento rappresenta un termine del vocabolario e assume il valore *vero* se la parola appare nel documento, e il valore *falso* altrimenti. Una query per un motore di ricerca costruito secondo questo modello può essere espressa collegando i termini con operatori booleani (es. *termine1 AND termine2 OR termine3*) e la risposta ad essa corrisponde al sottoinsieme di documenti della collezione la cui rappresentazione vettoriale soddisfa le condizioni espresse nella formula logica della query. Ogni documento sarà dunque semplicemente rilevante o non rilevante per una query e non vi sarà associata una misura di rilevanza; per questo motivo, i risultati di una ricerca non vengono restituiti secondo un ordinamento significativo.

Pur essendo un modello che consente di esprimere interrogazioni con precisione, la forma logica delle query costituisce sia un ostacolo sintattico per l'utente, che deve saper esprimere il proprio bisogno informativo in questa forma, sia un vincolo spesso troppo – o troppo poco – stringente nel processo di individuazione dei risultati rilevanti, in quanto i documenti che corrispondono precisamente alla query possono essere molto pochi, quando la query è molto specifica, o troppi, se comprende molti casi ed è espressa tramite un inseme condizioni disgiunte.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Figura 1.3 Rappresentazione di un insieme di documenti (sulle colonne) secondo il modello Booleano (termini sulle righe)

1.2.4 Term Frequency e Inverse Document Frequency

Per arricchire la rappresentazione di un documento utilizzata nel modello booleano si può tenere conto di due parametri relativi ad ogni termine del vocabolario: la frequenza con cui ogni termine appare in ciascun documento e la rarità di un termine rispetto all'intera collezione. L'idea alla base dell'introduzione di queste misure è che un documento sarà più rilevante rispetto ad una query se uno o più termini della query sono molto frequenti nel documento, oppure se uno o più termini che si trovano nella query sono presenti solo in quel documento o in pochi altri. La frequenza di una parola in un documento è detta Term Frequency [5]; matematicamente essa è definita come la frequenza assoluta di occorrenza del termine nel documento. Generalmente, questa misura non viene direttamente impiegata per calcolare la rilevanza di un documento rispetto a una query, poiché porterebbe a far aumentare il peso di un termine in un documento in modo spropositato, facendo risultare il documento di una rilevanza rispetto alla query maggiore di quella reale (un documento in cui un termine appaia 10 volte non dovrebbe essere 10 volte più significativo di uno in cui esso appare una volta sola). Per questa ragione, si impiega la funzione logaritmo per ridurre l'impatto della crescita del Term Frequency

sulla funzione di scoring; l'operazione applicata è detta *sublinear tf-scaling* ed è definita matematicamente come segue:

$$sub\_tfs(t,d) = \begin{cases} 1 + log_{10} tf_{t,d}, & tf_{t,d} > 0 \\ 0, & tf_{t,d} \leq 0' \end{cases}$$

dove t è un generico termine e d è un generico documento considerato.

La rarità di un termine in una collezione è definita matematicamente sulla base del concetto di Document Frequency di un termine, ossia il numero di documenti in cui il termine appare. Questa misura è, ovviamente, inversamente proporzionale alla rarità di un termine; la rarità viene quindi misurata da una misura chiamata Inverse Document Frequency [6], che è definita matematicamente come:

$$idf_t = log(\frac{N}{df}),$$

dove N è il numero di documenti nella collezione e df è definito come il Document Frequency del termine t.

Ad ogni termine in un documento può essere associato, dunque, un peso che è direttamente proporzionale sia alla frequenza del termine nel documento stesso, sia alla rarità del termine nella collezione; esso è il prodotto tra Term Frequency e Inverse Document Frequency del termine, ed è chiamato peso *tf-idf* del termine [7]:

$$w_{t,d} = sub\_tfs(w,d) \cdot idf_t$$

1.2.5 Il Vector Space Model

Il Modello a Spazio Vettoriale [8] è una tecnica di rappresentazione dei testi nella quale documenti e query sono rappresentati come vettori in uno stesso spazio multidimensionale, nel quale ogni dimensione rappresenta un

termine nel vocabolario della collezione. I testi sono quindi rappresentati da entità matematiche di tipo diverso rispetto al modello booleano, e di queste entità vengono sfruttate le proprietà per il calcolo della similarità tra query e documenti. Il modello integra, inoltre, le misure presentate nel paragrafo precedente: infatti, l'i-esimo elemento di un vettore che rappresenta un documento (o una query) ha come valore il peso associato al termine corrispondente rispetto a quel documento, calcolato in funzione della sua Term Frequency ed Inverse Document Frequency come descritto nel paragrafo precedente.

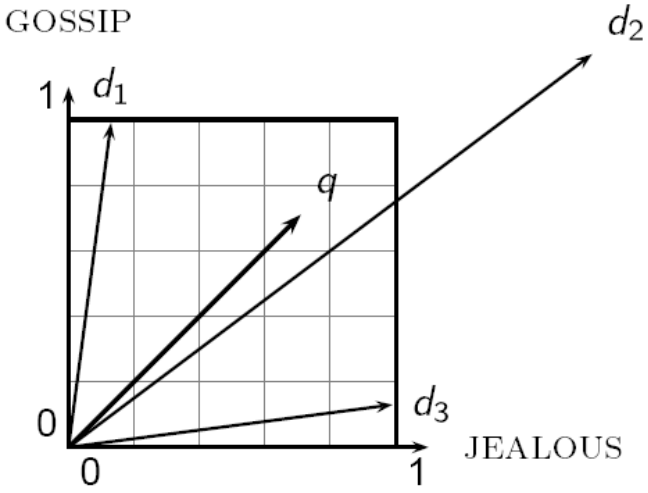

Figura 1.4 Rappresentazione dei documenti come vettori di lunghezza non normalizzata

Nella rappresentazione secondo il Modello a Spazio Vettoriale è possibile comparare documenti e query utilizzando operazioni vettoriali. Una prima possibilità per calcolare la similarità tra due documenti è considerare la loro distanza Euclidea nello spazio multidimensionale. Questa è definita come segue:

$$d(\vec{q},\vec{p}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2},$$

dove q_i e p_i sono gli elementi in posizione i-esima dei generici vettori $\vec{q}$ e $\vec{p}$. Usando questa metrica, però, due documenti, contenenti gli stessi termini

ma di dimensioni diverse a causa della ripetizione di alcuni termini in uno di essi, saranno considerati diversi perché distanti nello spazio. È necessario, quindi, compiere una normalizzazione della dimensione dei vettori. Questa operazione è compresa nel calcolo della misura di similarità del coseno, che valuta la distanza tra documenti sulla base dell'angolo compreso tra i vettori che li rappresentano. La metrica è definita come segue:

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \bullet \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sum_{i=1}^{|V|} q_i^2 \sum_{i=1}^{|V|} d_i^2}.$$

La normalizzazione rispetto alla lunghezza consiste nel dividere il prodotto interno tra i vettori per le loro norme.

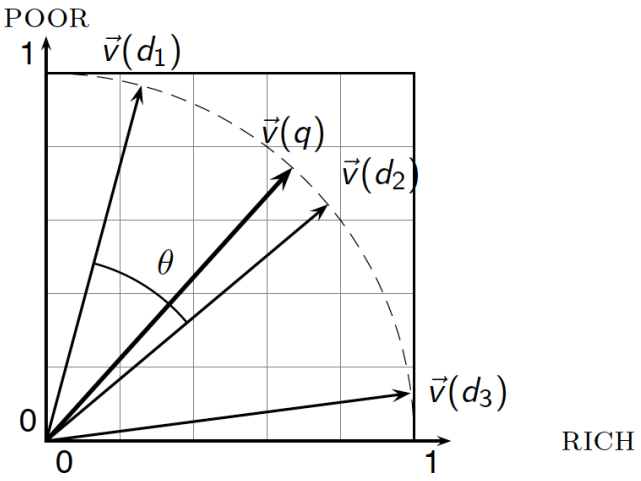

Figura 1.5 Rappresentazione della distanza tra vettori normalizzati basata sull'angolo compreso tra essi

Il Vector Space Model presenta comunque alcune problematiche:

1. come in tutte le tecniche di rappresentazione basate sul modello Bag of Words, la posizione dei termini non viene presa in considerazione;
2. documenti molto lunghi possono essere penalizzati nel processo di ritrovamento poiché tenderanno ad avere score di similarità più bassi a causa dell'operazione di normalizzazione delle lunghezze;

3. poiché viene usato il prodotto interno per confrontare i vettori, il modello può essere penalizzato dalla sparsità dei dati.

Inoltre, il modello a spazio vettoriale non può rilevare la similarità tra due documenti che abbiano un contenuto semanticamente simile ma contengano parole differenti, poiché la rappresentazione dei documenti è basata solo sulla forma delle parole che contengono e non sul loro significato.

1.3 Modelli Semantici Distribuzionali

Rappresentare il testo tenendo conto del significato delle parole ha un impatto significativo sulle operazioni di ritrovamento di documenti rilevanti rispetto a una query. Avere a disposizione una rappresentazione del significato delle parole usate in una collezione di documenti permette di calcolare la similarità tra una query e un documento non soltanto in base alle parole della query che occorono nel documento, ma anche considerando le parole in esso che hanno un significato simile a quelle date in input.

Per creare una rappresentazione del significato delle parole si utilizzano spesso modelli matematici che le rappresentano come vettori in uno spazio multidimensionale, nel quale quelle che hanno un significato simile sono rappresentate tramite vettori "vicini" tra loro; lo spazio multidimensionale viene detto Word Space o Semantic Space. In questo spazio si può misurare la similarità tra concetti calcolando il coseno dell'angolo compreso tra essi. L'assunzione fondamentale su cui si basano tutti i modelli semantici distribuzionali è che i termini che occorrono frequentemente in contesti simili – e raramente in contesti diversi – hanno un significato simile [9].

Esistono diversi metodi per creare lo spazio vettoriale nel quale rappresentare il significato delle parole; di seguito è presentata la tecnica chiamata Random Indexing.

1.3.1 Random Indexing

Il Random Indexing [11] è una delle tecniche di creazione di un Word Space su una collezione di documenti. Essa ha il vantaggio di generare una rappresentazione del significato delle parole senza utilizzare fonti di conoscenza esterne (come dizionari o ontologie) e in modo incrementale, ossia riuscendo ad aggiornare la rappresentazione in modo semplice all'aggiunta di nuovi documenti nella collezione; l'unica operazione di preprocessing del testo necessaria per costruire lo spazio è la tokenizzazione.

La tecnica consiste in due passi principali: il primo consiste nell'assegnare a ogni termine del vocabolario di una collezione un vettore ternario e sparso generato casualmente, ossia un vettore i cui elementi possono assumere i valori -1, 0 e 1, nel quale la maggior parte degli elementi hanno valore 0 e nel quale le posizioni degli elementi non nulli sono scelte in modo casuale. Il secondo passo consiste nel generare un "vettore contesto" per ogni termine nel vocabolario sommando i vettori precedentemente associati alle parole che occorrono frequentemente insieme ad esso nella collezione, in base ad una soglia di vicinanza prefissata. Il vettore contesto di un termine è calcolato secondo la formula seguente:

$$\overrightarrow{cv_i} = \sum_{d \in C} \sum_j \vec{r_j}, \ -c < j < c, i \neq j,$$

dove C è la collezione di documenti, c è la soglia di vicinanza dei termini prefissata che delimita il contesto di ogni parola e $\vec{r_j}$ è il vettore casuale

assegnato ad ogni parola presente nel contesto. Il vettore contesto di un termine è di fatto la rappresentazione del suo significato nello spazio multidimensionale, ed è un'implementazione algebrica del principio alla base dei modelli semantici distribuzionali enunciato in precedenza.

La creazione di un vettore contesto per ogni parola corrisponde formalmente ad un'operazione di riduzione della dimensionalità della matrice di co-occorrenza dei termini in una collezione. Questa operazione genera uno spazio di dimensionalità ridotta nel quale le distanze tra i vettori che rappresentano le parole restano proporzionali alle distanze che le parole avevano nello spazio orignale [10].

1.3.2 Motori di ricerca semantici

Usando la tecnica di Random Indexing è possibile rappresentare termini e documenti nello stesso spazio vettoriale multidimensionale. Si può generare, infatti, una rappresentazione vettoriale di un documento tramite la somma pesata di tutti i vettori che rappresentano le parole che contiene, usando come pesi gli indici di Inverse Document Frequency associati ad ogni parola; in questo modo è aumentata la rilevanza nella rappresentazione di un documento dei termini che sono meno comuni nella collezione.

Questa rappresentazione vettoriale di un documento nello stesso spazio in cui sono rappresentati i termin permette di calcolare la similarità semantica tra un termine e un documento o tra due documenti, oltre che tra due termini. Questa possibilità è alla base del funzionamento di un motore di ricerca semantico, ossia che tiene conto del significato dei termini. Se è possibile rappresentare un documento nello spazio vettoriale, infatti, si può rappresentare anche la query di un utente come il vettore somma dei

termini presenti in essa. Si può, dunque, calcolare la rilevanza dei documenti presenti nella collezione rispetto alla query come similarità del coseno tra il vettore che rappresenta la query e quelli che rappresentano i documenti; si può poi, come di consueto, ordinare i documenti in base alla rilevanza rispetto alla query, mostrando i primi *k* risultati all'utente.

2. Il progetto TALIA

Il progetto TALIA – acronimo di "Territorial Appropriation of Leading-edge Innovation Actions" – è un progetto del quale è capofila la Regione Puglia, finanziato dalla Commissione Europea nell'ambito di una call del programma Interreg-Med. Gli obiettivi più ampi della call erano la promozione dello sviluppo sostenibile supportato dalla tecnologia nell'area del Mediterraneo, e, più specificatamente, l'aumento della capacità di comunicazione e cooperazione tra gli attori principali nei più importanti settori socioeconomici nell'UE.

Il programma Interreg-Med comprende diverse decine di progetti denominati "verticali", portati avanti da imprese, enti pubblici e associazioni in diverse nazioni dell'area del mediterraneo. Questi progetti sono raggruppati, sulla base dei temi che affrontano, in nove gruppi chiamati "Community": Blue Growth, Green Growth, Social and Creative, Efficient Buildings, Renewable Energy, Urban Transports, Sustainable Tourism, Biodiversity Protection, Governance. A loro volta, le Community sono raggruppate in tre Assi, ciascuno dei quali comprende le Community con obiettivi comuni: Innovation Axis, Low Carbon Economy Axis, Natural and Cultural Resources Axis.

2.1 Obiettivi del progetto e soluzione adottata

Il progetto TALIA è trasversale ai progetti verticali del programma Med che afferiscono alla Community Social and Creative. Esso è stato sviluppato partendo dall'assunzione che fosse necessario andare oltre l'analisi dei singoli progetti verticali ed esplorare le loro possibilità di scalabilità mettendo ciascuno in relazione con gli altri. Per scalabilità si intende la capacità dei risultati dei progetti di raggiungere un numero maggiore di beneficiari nel

tempo e nello spazio. Attraverso il confronto dei risultati ottenuti dai diversi progetti e delle prassi adottate in ciascuno di essi, in relazione al contesto territoriale in cui sono inseriti, il progetto mira al miglioramento e alla crescita di ciascuno, anche tramite l'instaurazione di partnership e contaminazioni tra progetti simili, ma potenzialmente svolti in due luoghi distanti nell'area del mediterraneo.

Per raggiungere questi obiettivi, nell'ambito del progetto TALIA si è scelto di sviluppare un sistema di supporto decisionale basato sull'analisi intelligente del contenuto testuale dei documenti scritti in ogni progetto finanziato nell'ambito del programma Interreg-Med. Il sistema di supporto decisionale è basato sul Semantic Framework, una piattaforma per l'indicizzazione e l'analisi semantica dei documenti. Esso ha come utenti potenziali diversi tipi di soggetti coinvolti nell'ambito del programma Med:

- in primo luogo, i policy maker europei, chiamati a decidere l'indirizzo da prendere nelle call che stanzieranno in futuro i fondi comunitari destinati a quest'area di sviluppo; per questi, può essere rilevante tener conto dell'analisi automatica dei documenti prodotti nell'ambito di una o più Community del programma Med per comprendere, ad esempio, come queste abbiano operato, quali siano i temi principali intorno ai quali si è concentrata l'azione dei Partner di ogni progetto, e come diversi partner e progetti possano in futuro cooperare tra loro.

- In secondo luogo, i partner di ciascun progetto verticale possono beneficiare dai servizi di analisi automatica dei documenti per confrontare il proprio operato con quello svolto da altri soggetti in progetti

con obiettivi simili ai propri, per poter migliorare il proprio operato o valutare la possibilità di stabilire una collaborazione con essi.

- Infine, il sistema di supporto decisionale può supportare gli stakeholder di ciascun progetto, che possono essere enti pubblici, privati o singoli cittadini, a valutare l'andamento e l'impatto dei progetti operanti in aree di loro interesse specifico; ciò permette, dunque, al progetto TALIA di avere una maggiore capacità di coinvolgimento degli stakeholders sui territori, aumentando anche le potenzialità di ciascun progetto di innescare processi di crescita socioeconomica nelle aree in cui esso opera.

In particolare, il progetto TALIA, grazie al sistema di supporto alle decisioni basato sul Semantic Framework, mira a costruire e sviluppare la comunità Social and Creative del programma Interreg-Med; essa ha l'obiettivo di promuovere cluster di innovazione distribuiti tra le nazioni dell'area del mediterraneo fornendo strumenti che permettono la connessione di progetti trans-nazionali del programma Med con le comunità locali, a partire dalle regioni dei partner di ogni progetto.

2.2 Semantic Framework: architettura e funzionalità

Lo scopo del Semantic Framework è quello di rendere accessibile la conoscenza e le informazioni presenti nei deliverable di progetto prodotti nei progetti del programma MED. Il Framework si può considerare la parte di back-end del sistema di supporto alle decisioni sviluppato nel progetto TALIA, poiché offre le funzionalità per l'estrazione di informazioni che verranno poi utilizzate per soddisfare i bisogni informativi degli utenti.

Il Framework consente di gestire collezioni di documenti e di applicare ad ogni documento le operazioni di elaborazione del linguaggio naturale e di rappresentazione vettoriale e semantica descritte nel capitolo precedente. Esso è stato sviluppato secondo un'architettura client-server che permette l'accesso ai risultati del processo di elaborazione dei documenti tramite Web API sviluppate secondo il protocollo REST.

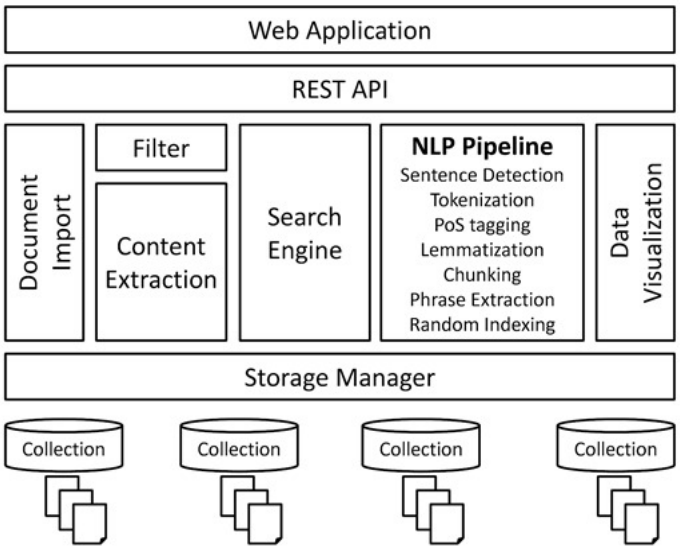

Figura 2.1 Architettura del Semantic Framework

2.2.1 Gestione delle Collezioni

La componente alla base del Semantic Framework, chiamata Storage Manager, è quella che consente di creare e gestire collezioni di documenti, ed è stata implementata grazie alla nota libreria Java Apache Lucene per la creazione di motori di ricerca. Nell'ambito dell'applicazione del Framework al progetto TALIA è stata creata una collezione di documenti per ogni Community relativa al programma Interreg-Med, e ciascuna di esse contiene tutti i deliverable di progetto che sono stati scritti nell'ambito di tutti progetti afferenti a ciascuna Community.

I deliverable di ogni progetto sono stati recuperati attraverso l'ausilio di un servizio software esterno al Semantic Framework che ha effettuato il crawling e lo scraping delle pagine web ufficiali del programma Med, scaricando i file dei deliverable e i metadati relativi a ogni progetto codificandoli in formato json. Il caricamento dei documenti in ciascuna collezione è avvenuto in seguito a un processo di selezione dei documenti significativi per ogni Community, al fine di ottenere collezioni contenenti esclusivamente documenti in lingua inglese o italiana, con un contenuto informativo utile, e privi di immagini, condizioni necessarie affiché il processo automatico di estrazione e analisi del testo desse risultati corretti, comprensibili e significativi. Per facilitare il processo di selezione dei documenti e di caricamento di quelli significativi nella relativa collezione è stato sviluppato un ulteriore servizio web esterno alla piattaforma, che permette di visualizzare i documenti estratti tramite il processo di crawling, selezionare manualmente quelli di interesse dell'utente e caricarli automaticamente nel Semantic Framework attraverso i servizi REST di gestione delle collezioni.

Attraverso i servizi REST esposti dal Framework, ogni collezione può essere creata specificandone il nome e la lingua – informazione necessaria, quest'ultima, durante il processo di NLP. Oltre a poter creare nuove collezioni vuote e poi aggiungervi documenti, è possibile generarne di nuove attraverso la fusione di due o più collezioni già esistenti. Sono state create in questo modo, ad esempio, le collezioni di documenti relative agli Assi tematici del programma MED. Per ogni collezione possono essere aggiunti o rimossi documenti; al momento del caricamento in una collezione, ogni documento può essere strutturato in sezioni (come "titolo", "autore", "corpo", ecc.), grazie alle funzionalità messe a disposizione dalla

libreria Lucene. Un documento può essere aggiunto in formato CSV o PDF; in entrambi i casi, sarà conservato all'interno della piattaforma come un file contenente solo testo, che, nel caso dei file PDF, viene estratto automaticamente tramite la libreria Apache Tika.

2.2.2 Motori di ricerca

Per poter estrarre informazioni e conoscenza dai documenti presenti nelle collezioni, su ogni documento è possibile eseguire tutti i passi della pipeline di NLP descritti nel paragrafo 1.1. Questi generano all'interno della piattaforma i file contenenti i risultati del processo di elaborazione del linguaggio naturale; per ogni documento viene creato un file conenente i token estratti da esso, il lemma corrispondente, il POS Tag associato ad essi e la parte che occupano nel "chunk" di cui fanno parte. Inoltre, viene creato un file contenenti le key phrases di ogni collezione.

Sulla base dei risultati delle operazioni di elaborazione del testo, la piattaforma è in grado di compiere operazioni di indicizzazione dei documenti basate sul modello a spazio vettoriale, presentato nel paragrafo 1.2. Inoltre, è possibile creare una rappresentazione semantica dei testi attraverso la tecnica di creazione di un modello semantico distribuzionale chiamata Random Indexing, presentata nel paragrafo 1.3.1. In seguito alle operazioni di preprocessing, indicizzazione e creazione di un Semantic Space associato ai documenti è possibile creare e associare a ciascuna collezione un motore di ricerca classico o semantico, a seconda delle operazioni di elaborazione dei documenti svolte.

2.2.3 Gestione dei metadati

Nell'ambito del progetto TALIA, oltre a beneficiare della conoscenza estratta dai documenti, è risultato utile sfruttare le informazioni strutturate disponibili relative ad ogni community e progetto.

Sono stati confrontati due approcci alla gestione dei metadati: il primo si basa sulla funzionalità della piattaforma Apache Lucene che consente di suddividere un documento in campi (es. titolo, sottotitolo, autore, abstract, corpo, ecc.), i quali conterrebbero le informazioni struttrate relative al documento. Il secondo impiega una base di dati relazionale e un modulo del Semantic Framework di interfaccia ad essa per la gestione dei dati strutturati. Sebbene la prima soluzione fosse già integrata nel Semantic Framework (poiché esso si basa su Apache Lucene), essa presentava anche numerosi problemi nella gestione della ridondanza e dell'integrità nei dati; pertanto, si è scelto di sviluppare un modulo estensivo del Semantic Framework per l'interfaccia ad una base di dati relazionale, che consentisse naturalmente di gestire i problemi citati.

Tale modulo è in grado di rendere disponibili informazioni dettagliate riguardo ogni documento presente nelle collezioni gestite nel Framework, e soprattutto relativamente al contesto in cui esso è inserito nell'ambito del programma Med. Ad esempio, oltre a quelle relative ad ogni singolo documento, sono disponibili informazioni dettagliate relative al progetto verticale cui ogni deliverable afferisce, ai partner e agli stakeholder coinvolti in esso. La disponibilità di queste informazioni aggiuntive sul dominio in forma strutturata è stata utile ad arricchire ed ampliare i risultati ottenuti utilizzando i servizi del Sistema di Supporto alle Decisioni, affiancando la potenzialità informativa offerta dalle funzionalità di analisi semantica del testo.

2.3 Servizi del Sistema di Supporto alle Decisioni

Nell'ambito del progetto TALIA, le funzionalità del Semantic Framework sono state utilizzate per l'implementazione dei due principali servizi del

sistema di supporto alle decisioni: un motore di ricerca semantico e una matrice di correlazione tra concetti.

I due servizi hanno l'obiettivo di contribuire al miglioramento del coordinamento e della cooperazione tra gli attori principali della Community Social and Creative facilitando la scoperta di informazioni utili e connessioni esistenti tra le azioni concrete svolte dai diversi attori, supportando le attività degli utenti come descritto nel paragrafo 2.1.

2.3.1 Motore di ricerca semantico

Il motore di ricerca semantico ha l'obiettivo di supportare gli utenti del sistema di supporto alle decisioni nel ritrovare i deliverable di progetto maggiormente significativi rispetto ad un tema più o meno complesso, espresso tramite una query. Il motore di ricerca offre la possibilità esplorare le collezioni di documenti relativi a una Community o ad un Asse del programma Interreg-Med.

Inoltre, i risultati di una ricerca sono ulteriormente arricchiti sia grazie alle funzionalità di rappresentazione semantica del testo rese disponibili dal Semantic Framework, che grazie alla disponibilità di metadati relativi ai documenti e al dominio. L'utente riceve, infatti, in risposta ad una query, l'elenco dei documenti più significativi rispetto ad essa, un elenco dei concetti maggiormente correlati a quello cercato e un insieme di metadati che forniscono informazioni più dettagliate relative a ciascun documento e al contesto in cui è inserito nel Programma Med. I concetti correlati alla query che vengono restituiti insieme ai risultati sono, inoltre, punti di partenza per una nuova ricerca nella collezione selezionata.

Un'altra funzionalità disponibile nel servizio di ricerca semantica è il riassunto automatico dei documenti, abilitato dalla rappresentazione

semantica dei testi [14]. I documenti risultanti da una ricerca sono spesso molti e molto lunghi; per supportare l'utente nell'analisi dei risultati è stato reso disponibile un servizio che, dato in input un documento e una misura di ampiezza del riassunto che si vuole ottenere (espresso come numero di frasi o in percentuale rispetto al testo originale), restituisce un testo che contiene le informazioni più significative del documento originale, dando all'utente la possibilità di avere un'idea più precisa degli argomenti trattato in esso.

2.3.2 Matrice di correlazione tra concetti

La matrice di correlazione dà all'utente del sistema di supporto alle decisioni la possibilità di esplorare in modo interattivo il contenuto informativo dei documenti di una collezione, beneficiando dell'analisi semantica dei testi svolta grazie al Semantic Framework.

L'esplorazione di una collezione consiste nel poter visualizzare, data una coppia di concetti, un elenco contenente i concetti maggiormente correlati ad essi all'interno della collezione considerata. Grazie alla presentazione della funzionalità di correlazione sottoforma di matrice è possibile compiere questa analisi su diverse coppie di concetti contemporaneamente. Assegnando un concetto ad ogni riga e ad ogni colonna, infatti, ciascuna cella della matrice corrisponderà ai concetti associati alla riga e alla colonna che identificano la cella; essa sarà quindi riempita con i concetti della collezione maggiormente correlati ad essi. I concetti corrispondenti alle dimensioni della matrice possono essere, inoltre, sia concetti presenti nei documenti della collezione, sia definiti dal creatore della matrice a partire da quelli già esistenti.

La creazione delle liste di concetti che riempiono le celle consiste, naturalmente, nella fusione delle liste di concetti maggiormente correlati a ciascun concetto della coppia associata ad ogni cella. L'algoritmo attualmente implementato per la fusione delle liste di concetti è chiamato CombSUM; esso è stato studiato nel campo dell'Information Retrieval come metodo di fusione dei risultati provenienti da due diversi algoritmi di ranking dei risultati di una ricerca. Il suo semplice funzionamento è uno dei motivi per cui si è scelto di implementarlo: dati due ranking nei quali ad ogni documento è associato un peso, l'algoritmo genera un nuovo ranking che comprende tutti i documenti presenti nei precedenti, nel quale ogni documento avrà come peso la somma dei pesi che aveva nei due ranking iniziali (se un documento non appare in uno dei due ranking avrà peso nullo relativamente ad esso).

La matrice di correlazione tra concetti, servizio più elaborato tra quelli messi a disposizione dal sistema di supporto alle decisioni, permette all'utente di compiere analisi più dettagliate e più flessibili rispetto al servizio di ricerca semantica. Ciascuna matrice, infatti, definita dalle sue dimensioni e dai concetti associati ad esse, può essere vista come una specifica analisi compiuta sul dominio rappresentato dalla collezione (e Community) sulla base della quale vengono calcolati i concetti correlati. Ogni analisi può essere creata da un utente oppure progettata da un esperto; questa può essere quindi utilizzata da altri utenti, sia così com'è stata creata, che modificando le definizioni dei concetti in base alle proprie esigenze. I concetti maggiormente correlati a quelli definiti, mostrati nelle celle della matrice, possono essere poi interpretati in diversi modi, disponendo di un'adeguata conoscenza del dominio e dello strumento usato. Infatti, essi possono essere considerati sia come una "fotografia"

dello stato attuale dei progetti operanti nella Community scelta, sia interpretati con il fine di prescrivere possibili azioni da compiere per cambiare i risultati ottenuti, e, di conseguenza, la situazione reale che si ritiene il servizio stia rappresentando.

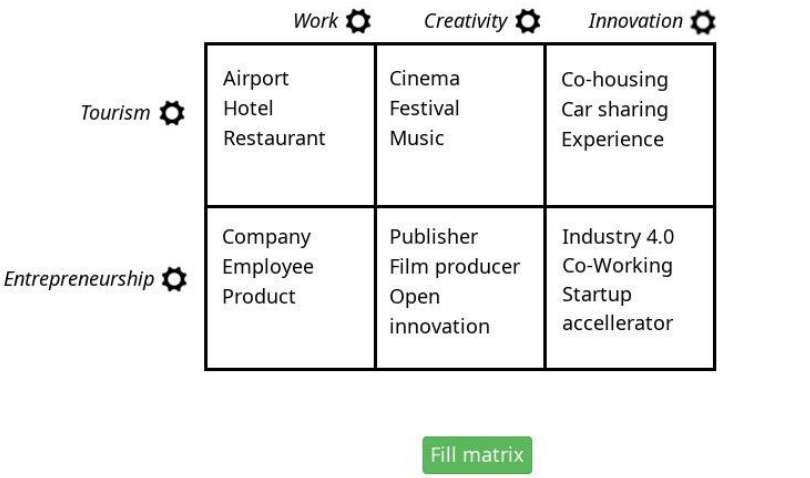

Figura 2.2 Una immagine di mockup della matrice di correlazione tra concetti

3. Dati strutturati e servizi RESTful

L'implementazione dei casi d'uso ha richiesto l'utilizzo delle funzionalità offerte dal Semantic Framework esposte tramite servizi RESTful. Ogni caso d'uso ha richiesto l'utilizzo, l'implementazione e il miglioramento di numerosi servizi specifici, sia al fine di realizzare le funzionalità richieste, che sia per aumentare l'efficacia dei servizi già esistenti.

3.1 Introduzione alle architetture RESTful

Lo stile architetturale REST è una tecnica di sviluppo di servizi Web concepita nel 2000 da Roy Fielding [12], allora studente di dottorato all'Università della California a Irvine.

Le architetture REST (ossia REpresentational State Transfert) impiegano il protocollo HTTP per trasferire tutte le informazioni in una comunicazione tra host client e server. Nel protocollo HTTP ogni entità che si può richiedere ad un host server è considerata una "risorsa", identificabile e localizzabile univocamente tramite il suo indirizzo URL [13]. Ogni risorsa, in un dato momento della comunicazione, si trova in uno "stato"; tramite un servizio REST si può, quindi, o richiedere una rappresentazione dello stato di una risorsa, oppure trasmettere una rappresentazione dello stato che si vorrebbe la risorsa raggiungesse, per richiederne la modifica. Dai concetti di risorsa, stato e rappresentazione di uno stato deriva, dunque, l'acronimo REpresentational State Transfert.

Un servizio web sviluppato secondo questo stile architetturale, dunque, è accessibile a un computer host tramite una semplice richiesta HTTP all'host server che espone il servizio. A seconda che il servizio richiamato abbia lo scopo di modificare o ottenere informazioni presenti sull'host server, la richiesta HTTP dev'essere formulata tramite i metodi HTTP progettati per

ciascuno scopo particolare; ad esempio, il metodo GET è specifico per la richiesta di una risorsa senza modifica, mentre i metodi POST e PUT sono specifici per le modifiche delle risorse.

I vantaggi dell'uso di un'architettura REST sono la sua grande flessibilità, l'indipendenza tra client e server nella comunicazione e la leggerezza dei messaggi scambiati, grazie alla quale si ottengono ottime performance.

3.2 Metadati strutturati

Il dominio del progetto TALIA, coinvolgendo numerosi progetti verticali ulteriormente raggruppati su vari livelli, offre una ricchezza di informazioni che possono essere rese disponibili proficuamente ai policy makers e agli altri utenti del Sistema di Supporto alle Secisioni, integrandoli nei servizi offerti dal Sistema stesso.

Come accennato nel paragrafo 2.2.3, sono state confrontate due soluzioni alternative per la gestione dei dati strutturati relativi al dominio.

La prima soluzione consiste nel gestire le informazioni strutturate, reperite durante il processo di crawling che ha permesso di ottenere anche i deliverable dei progetti del programma Med, impiegando le funzionalità offerte dalla libreria Apache Lucene. In particolare, la libreria per la creazione e gestione di motori di ricerca dà all'utente la possibilità di suddividere ciascun documento il più sezioni (o campi), al fine di permettere di compiere una ricerca mirata sul contenuto di alcuni campi scelti (ad esempio, ricerca sui titoli o sugli abstract dei documenti). Le informazioni più importanti relative al dominio e a ciascun documento possono essere quindi memorizzate all'interno del Semantic Framework come campi associati ai documenti.

Questa soluzione presenta alcuni vantaggi:

- In primo luogo, poiché all'interno del Semantic Framework la gestione dei documenti è basata sulla libreria Apache Lucene, questa soluzione può essere immediatamente implementata poiché non vi è necessità di alcuna operazione di sviluppo o integrazione;
- In secondo luogo, ad ogni documento vengono associate naturalmente le informazioni relative ad esso, pertanto la consultazione delle informazioni relative ad ogni singolo documento risulta efficiente ed efficace;
- In terzo luogo, considerando le esigenze implementative e di integrazione con i servizi del Semantic Framework, poiché ad ogni documento sono associate le informazioni ad esso relative, è immediato restituire come risultato di una ricerca nelle collezioni di documenti, assieme ad ogni documento, anche le informazioni relative ad esso.

Tuttavia, il dominio del progetto TALIA – che coinvolge diversi progetti verticali del programma Med – comprende diverse entità relativamente alle quali sono disponibili informazioni, non soltanto i documenti. Tra le entità del dominio, tra l'altro, sono presenti relazioni che non è possibile rappresentare efficacemente utilizzando la funzionalità di strutturazione dei documenti della libreria Apache Lucene. Questa soluzione presenta, infatti, anche diverse pecche:

- In primo luogo, la quantità di informazioni memorizzabili relativamente a ciascun documento è limitata, soprattutto rispetto all'abbondanza di informazioni disponibili sul dominio dei progetti Med. Infatti, aggiungere troppi campi ad un documento fa sì che esso

occupi più spazio di memorizzazione del necessario, oltre a non rispettare il fine originale dei campi di un documento, poiché ben pochi campi che contengono metadati sono adatti a poter compiere una ricerca su di essi.

- In secondo luogo, molte informazioni sono ridondanti poiché ripetute in diversi documenti, nonostante esistano relazioni tra le entità del dominio tenendo conto delle quali si potrebbe semplificare la rappresentazione delle informazioni (ad esempio, ad un Progetto corrispondono diversi Deliverable, dunque le informazioni relative al progetto possono non essere ripetute in ogni deliverable relativo ad esso, il che invece accade memorizzando i metadati come campi del documento).

- In terzo luogo, i vincoli di integrità esistenti tra le informazioni del dominio sono difficilmente preservabili in un modello che le metta in relazione soltanto con il documento cui si riferiscono, e non le metta, invece, in relazione tra loro rispecchiando le connessioni realmente esistenti tra le entità. Risulta, dunque, molto difficile rilevare errori o incongruenze nei dati.

- Infine, non potendo codificare le relazioni esistenti tra le entità del dominio, non è possibile neanche compiere ricerche più sofisticate sui dati disponibili, prescindendo dall'associazione tra essi e un particolare documento della collezione.

L'esigenza di poter disporre dei dati in modo più flessibile è dovuta, in particolare, alla necessità di implementazione di servizi del sistema di supporto alle decisioni più sofisticati, come la visualizzazione geografica dei risultati di ricerca illustrata nel paragrafo 4.2.

Per evitare le pecche della prima soluzione illustrata, e mantenerne contemporaneamente i vantaggi, è stato integrato nel Semantic Framework un modulo di gestione dei metadati relativi ad ogni collezione. Esso collega il Semantic Framework ad una base di dati relazionale progettata per contenere e organizzare i dati relativi ai documenti presenti nelle collezioni, grazie ad una struttura che rispecchia le relazioni esistenti tra le entità coinvolte nel dominio.

I dati strutturati disponibili – chiamati "metadati", in quanto sono informazioni relative a documenti che già di per sé forniscono informazioni sui diversi progetti – sono stati reperiti tramite un processo di web scraping, e successivamente inseriti, tramite servizi RESTful, in una base di dati progettata secondo un modello Entità-Relazione del dominio.

Nei paragrafi successivi viene illustrato il processo di acquisizione dei metadati, il modello E-R progettato e i servizi RESTful sviluppati per la gestione dei metadati.

3.2.1 Sorgenti dei dati

Il processo di estrazione dei metadati dal sito web ufficiale di ogni progetto del programma Interreg-Med ha prodotto i seguenti file: un file in formato Json per ogni Community, ciascuno contenente informazioni relative a tutti i progetti afferenti alla Community associata; due file in formato Xlsx, contenenti informazioni relative all'intero programma Med.

Ogni file Json è incentrato sui deliverable. La componente principale della sua struttura è una lista di oggetti che rappresentano i deliverables pubblici di tutti i progetti afferenti alla Community cui il file si riferisce. Ogni oggetto comprende, tra le altre, informazioni relative al deliverable stesso, al progetto cui afferisce e ai partner coinvolti in esso.

Figura 3.1 Struttura del file Json relativo alla Community "Blue Growth"

Dei due file Xlsx generati dal processo di crawling e scraping, uno contiene informazioni relative a tutti i progetti verticali finanziati nell'ambito dell'intero programma Med, e l'altro contiene informazioni relative ad ogni partner coinvolto in essi.

Axis	Objective	Acronym	Project label	Operation summary	Lead Partner	Country	Postcode	Call for proposals	Start date	End date	Type of project	ERDF	IPA Funds	Amount of the project (ERDF +IPA + national counterpart)	Co financing rate	Community
1	1	+RESILIENT	Mediterranean Open REsouRCEs for Social Innovation of Socially Responsive ENTerprises	+RESILIENT puts together a 4-helix partnership of 8 MED countries to tackle the need for innovation conducive to increased socially-responsive competitiveness of SMEs & stimulate new jobs, especially for companies operating in the social economy. It aims to kickstart a process of policy change at regional	Veneto Region – Operational Unit for EU and State Relations	ITALY	30123	2nd call	01/02/2018	31/01/2022	Integrated project	2.667.650,20	119.100,13	3.278.529,80	85%	Social and Creative

Figura 3.2 Struttura del file Xlsx relativo ai Progetti

Sebbene i file Json ed Xlsx siano stati generati durante il medesimo processo di estrazione, essi non contengono le stesse informazioni. Ad esempio, poiché i file Json mettono in relazione ogni deliverable con il relativo progetto e i partner che vi partecipano, essi contengono dati ridondanti, sia relativamente al file stesso, sia in relazione ai file Xlsx riguardanti i progetti o i partner. Inoltre, i file Json contengono alcuni campi che non sono contenuti nei file Xlsx, e viceversa. Questa abbondanza di informazioni è stata analizzata e considerata accuratamente per progettare un modello Entità-Relazione del dominio che fosse il più completo possibile.

Per essere inseriti nel database tramite i servizi RESTful i file Xlsx sono stati convertiti nel formato Csv, più semplice da gestire.

3.2.2 Modello Entità-Relazione del dominio

Il modello E-R del dominio è stato progettato analizzando la struttura dei file di dati descritti in precedenza ed integrando ulteriore conoscenza del dominio resa disponibile da esperti. I prinicipali concetti rappresentati come entità nel modello sono: i Deliverable, i Progetti, i Partner di un progetto, gli Stakeholder di un progetto e le Community. In conformità con le caratteristiche delle relazioni realmente esistenti nel dominio tra queste entità, le relazioni nel modello sono state modellate tutte come relazioni uno-a-molti, eccetto quelle relative ai progetti e ai loro partner e stakeholder, che sono state modellate con cardinalità molti-a-molti (in quanto, ogni Progetto può coinvolgere uno o più Partner e ogni Partner può partecipare ad uno o più Progetti). Sono state infine inserite nel modello le entità necessarie a rappresentare le Keywords e i Targets associati ad ogni Deliverable e Stakeholder.

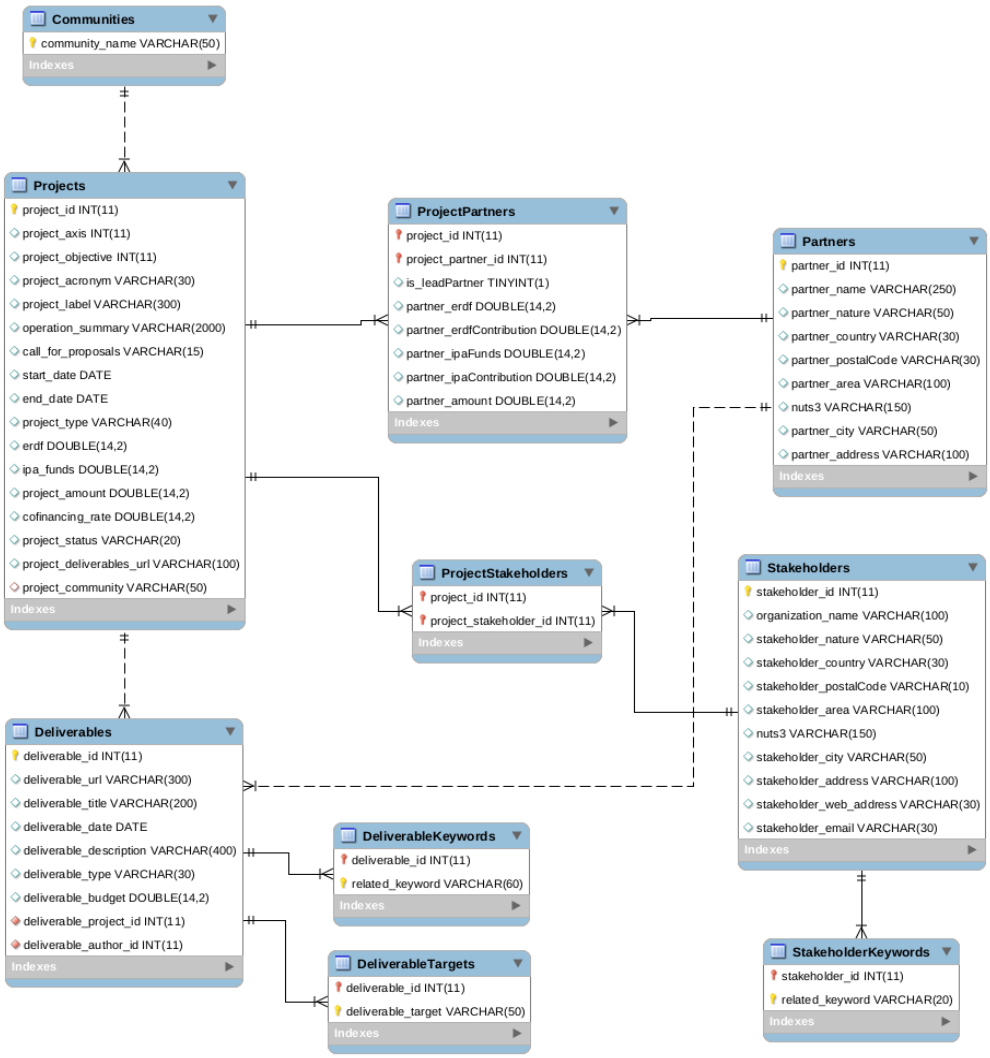

Figura 3.3 Struttura del modello Entità-Relazione

3.3 Servizi RESTful per la gestione dei metadati

Come accennato in precedenza, il Semantic Framework è stato ampliato per poter gestire metadati relativi alle collezioni e ai documenti presenti in esse e integrare così alle informazioni ottenute tramite l'analisi semantica dei documenti anche le informazioni strutturate disponibili sul dominio dei documenti considerati. Nel caso del progetto TALIA, il dominio è relativo alle Community del programma Interreg-Med.

La gestione dei metadati ha richiesto la progettazione di moduli software per l'acquisizione automatica dei dati dai file citati in precedenza, e il successivo inserimento nel database realizzato secondo la struttura del modello Entità-Relazione presentato nel paragrafo 3.2.2.

3.3.1 Gestire i dati provenienti dai file

Per l'aquisizione dei dati contenuti nei file Json e Csv sono state realizzate due classi che operano un processo di parsing su ciascun file secondo la sua struttura. Per svolgere in modo efficiente il parsing dei file Json sono state create delle classi innestate che rispecchiassero la struttura interna al file, in modo tale da poter riportare il contenuto presente in essi in nell'oggetto che rappresenta l'intero file Json, e che potesse essere poi manipolato dinamicamente. Il parsing dei file Csv, invece, è stato svolto semplicemente grazie alla loro struttura.

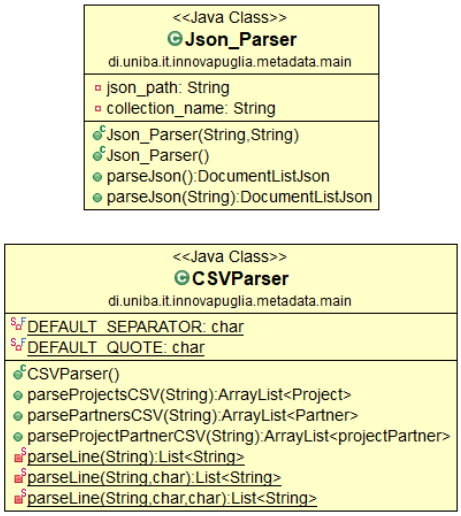

Figura 3.4 Classi per il parsing dei file Json e Csv

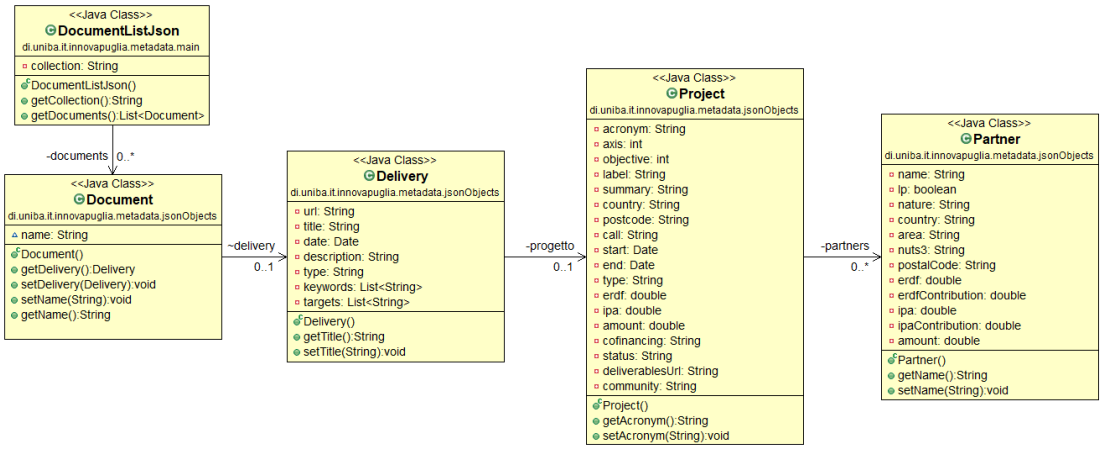

Figura 3.5 Package delle classi che rispecchiano la struttura dei file Json

3.3.2 Gestire i dati contenuti nella base di dati

In modo analogo a quanto fatto per i file Json, per manipolare i dati contenuti nel database all'interno del Semantic Framework, sono state progettate delle classi che rispecchiassero la struttura del modello Entità-Relazione sulla base del quale è stata progettata la base di dati. Questo ha consentito gestire semplicemente i risultati delle interrogazioni alla base di dati svolte tramite le API REST e il modulo presentato nel paragrafo seguente.

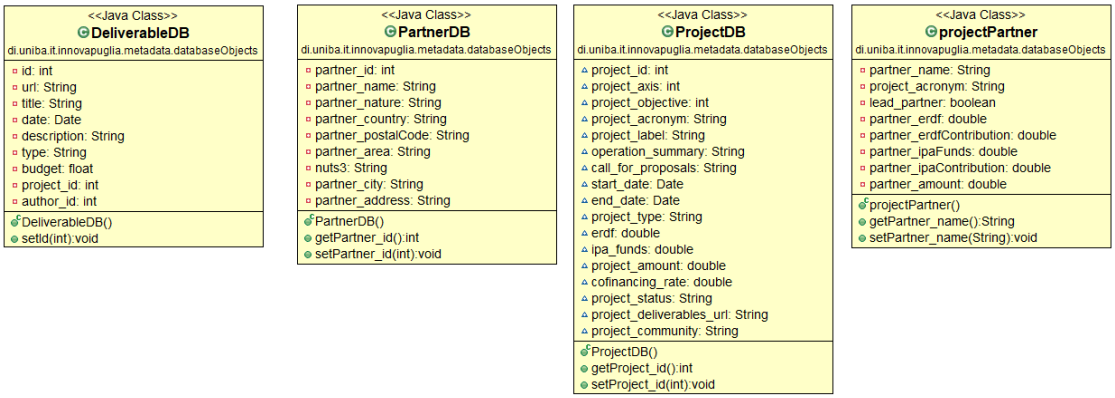

Figura 3.6 Package delle classi che rispecchiano la struttura del database

3.3.3 Gestire la connessione con la base di dati

La connessione al database è stata gestita tramite un'unica classe denominata *DatabaseInterface*, la quale, implementando il driver di connessione Java con database MySql, espone metodi che corrispondono alle query di inserimento dati e interrogazione del database utili per integrare i dati strutturati ai servizi del Sistema di Supporto alle Decisioni.

Per poter gestire efficacemente il caricamento dei dati nel database svolto tramite le API REST a partire dai file Json e Csv dati in input, è stata creata, inoltre, la classe *MultipleDBRequest*. Questa espone metodi per il caricamento di tutti i dati relativi ai progetti del programma Med, ai partner coinvolti in essi, o ad una specifica Community, che utilizzano, a loro volta, i metodi esposti dalla classe *DatabaseInterface* per compiere le singole query di inserimento.

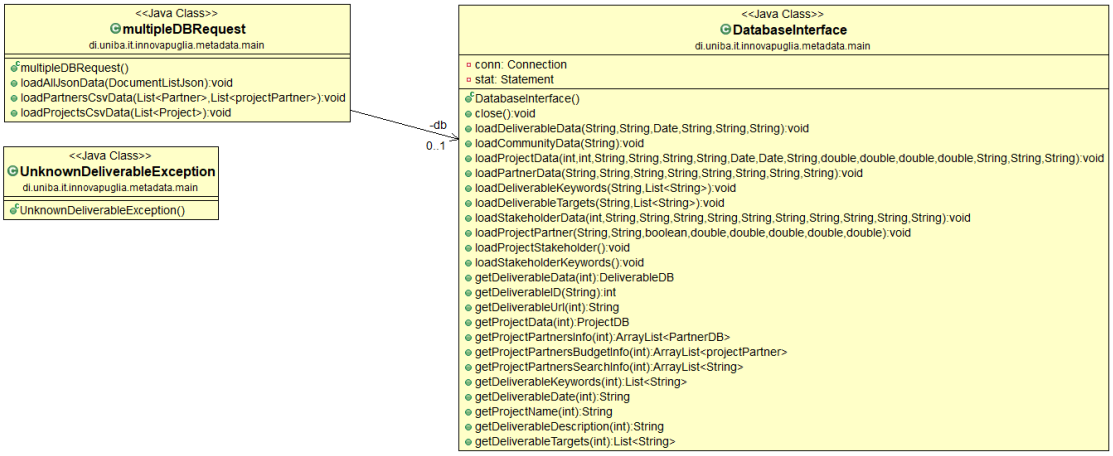

Figura 3.7 Package delle classi di interfaccia con il database

3.4 API REST di interfaccia al database

I servizi di gestione dei metadati relativi alle collezioni illustrati in precedenza sono stati resi disponibili all'utente tramite API REST, in conformità con l'architettura del Semantic Framework.

3.4.1 API REST per il popolamento del database

Come accennato nei paragrafi precedenti, tramite i metodi REST per il popolamento del database, è possibile, inviando da una stringa Json al server del Semantic Framework nel corpo della chiamata *PUT* del metodo apposito, inserire nel database tutti i dati contenuti nella stringa Json relativi alla Community (e, quindi, alla collezione) cui esso si riferisce.

Request	*PUT http://90.147.102.57:9001/ivp/v2/loadJsonData*
Body	{"collection":"Biodiversity Protection","documents":[{"name":"Confish_Deliverable_3.4.1_Genotype-by-sequencing_raw_genomic_data.pdf","delivery":{"url":"https://confish.interreg-med.eu/what-we-achieve/deliverable-library/detail/[...]","title":"Genotype-by-sequencing raw genomic data","date":"2018-09-30","description":"Raw sequencing data on Mediterranean populations of common octopus and read and blue shrimp","type":"Document","keywords":["climate change","biodiversity","coastal management","maritime issues","sustainable management of natural resources"],"progetto":{"acronym":"ConFish","axis":3,"objective":2,"label":"Connectivity among Mediterranean fishery stakeholders and scientists resolves connectivity of fishery populations","summary":"In this era of fast global change, [...]","country":"CROATIA","postcode":"10000","call":"1st call","start":"2016-10-31","end":"2018-07-30","type":"Studying","erdf":477357.03,"ipa":0.0,"amount":561596.51,"cofinancing":"0.85","status":"On going","deliverablesUrl":"http://confish.interreg-med.eu/what-we-achieve/deliverable-library/","partners":[{"name":"Faculty of Science, University of Zagreb","postalCode":"10000","erdf":143201.2,"erdfContribution":25270.8,"ipa":0.0,"ipaContribution":0.0,"amount":168472.0,"country":"CROATIA","lp":true,"area":"Kontinentalna Hrvatska","nature":"Higher education and research","nuts3":"Grad Zagreb"}]},"targets":["Higher education and research"]}}]}

Tabella 3.1 Esempio di chiamata al metodo per l'inserimento dei dati relativi a una Community

Analogamente, è possibile inviare nel corpo della chiamata al metodo apposito la stringa in formato Csv contenente i dati relativi ai Progetti o ai Partner di tutto il programma Interreg-Med.

Request	*PUT https://localhost:9001/ivp/v2/loadPartnersCsvData*
Body	`1;1;+ RESILIENT;1;Veneto Region - Operational Unit for EU and State Relations;Regional Public authority;ITALY;Veneto;Venezia;30123;339.043,75;59.831,25;0,00;0,00;398.875,00` `1;1;+ RESILIENT;0;AIX-MARSEILLE UNIVERSITY;Higher education and research;FRANCE;Provence-Alpes-Côte d'Azur;Bouches-du-Rhône;13284;222.976,25;39.348,75;0,00;0,00;262.325,00`

Tabella 3.2 Esempio di chiamata al metodo per l'inserimento dei dati relativi ai Partner dei Progetti del programma Interreg-Med

L'inserimento dei dati è possibile grazie al parsing delle stringhe nei due formati e al caricamento dei dati nel database tramite le classi illustrate in precedenza.

3.4.2 API REST per l'interrogazione del database

Al fine di fornire i dati necessari all'implementazione di servizi avanzati del Sistema di Supporto alle Decisioni, sono stati creati dei metodi REST tramite i quali è possibile ottenere le informazioni contenute nel database relative a singoli documenti (Deliverable), Partner o Progetti.

I metodi REST esposti non sono in rapporto uno ad uno con i metodi presenti nella classe *DatabaseInterface*, che compiono le interrogazioni al database tramite il driver MySql. Essi, infatti, pur utilizzando i metodi esposti dalla classe, spesso filtrano i risultati delle query esponendo di fatto metodi di accesso ai dati di granularità più fine rispetto a quelli esposti dalla classe di interfaccia diretta. Ciò è stato fatto per fornire all'utente risultati più specifici, adattati alle particolari informazioni che è necessario integrare nei servizi del Sistema di Supporto alle Decisioni.

Tutti i dati richiesti tramite API REST vengono restituiti dai metodi in formato Json (Tabella 3.3), per garantire l'interoperabilità con altri sistemi, come il servizio di creazione di mappe illustrato nel paragrafo successivo.

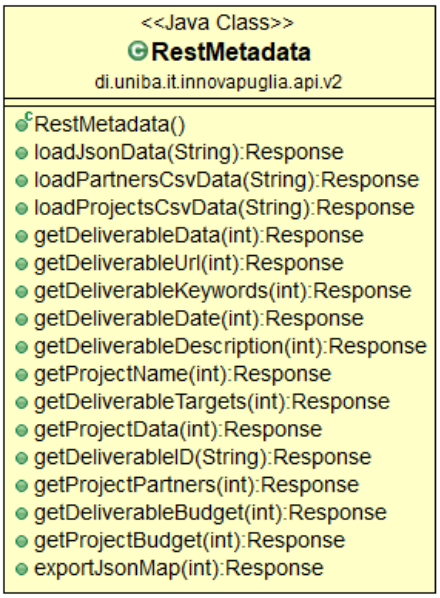

Figura 3.8 Classe che espone le API REST

Request	*GET http://90.147.102.57:9001/ivp/v2/getDeliverableData/1*
Answer	{"id":1,"url":"https://confish.interreg-med.eu/what-we-achieve/deliverable-library/detail/?tx_elibrary_pi1%5Blivrable%5D=5998&tx_elibrary_pi1%5Baction%5D=show&tx_elibrary_pi1%5Bcontroller%5D=Frontend%5CLivrable&cHash=816f19d7af2fd446073fb9190fc3c67b","title":"Confish_Deliverable_3.4.1_Genotype-by-sequencing_raw_genomic_data","date":1538265600000,"description":"Raw sequencing data on Mediterranean populations of common octopus and read and blue shrimp","type":"Document","budget":50514.27,"project_id":20,"author_id":186}

Tabella 3.3 Esempio di chiamata al metodo di interrogazione al database sulla tabella Deliverables

4. Informazioni strutturate nei servizi del Sistema di Supporto alle Decisioni

Come illustrato nel paragrafo 2.3, il Sistema di Supporto alle Decisioni sviluppato nell'ambito del progetto TALIA offre due servizi principali basati sulla rappresentazione semantica del testo generata dal Semantic Framework: un motore di ricerca semantico e una matrice di correlazione tra concetti. I dati strutturati riguardanti il dominio dei progetti Interreg-Med sono stati integrati principalmente nel servizio di ricerca semantica e, di riflesso, nel servizio di correlazione tra concetti.

La disponibilità di dati strutturati è stata, inoltre, fondamentale per la realizzazione di un servizio interno alla ricerca semantica che consente la visualizzazione geografica dei risultati di ricerca. Le mappe generate sono dinamiche, poiché sono generate in base ai risultati di ricerca ottenuti a partire da una query; esse impiegano i metadati per rappresentare graficamente diverse informazioni riguardanti i documenti presenti nel ranking. La ricchezza di informazioni, la flessibilità e la semplicità nell'accesso ad esse ha consentito di creare mappe tematiche di tipi diversi, che saranno illustrate nel paragrafo 4.3.

Tutti i servizi del sistema di supporto alle decisioni fin qui illustrati sono stati sviluppati come servizi Web, nei quali è stato naturale utilizzare le API REST descritte nel paragrafo precedente.

4.1 Informazioni strutturate per la ricerca semantica

L'integrazione dei dati strutturati nel servizio di ricerca semantica ha permesso di rendere maggiormente informativi i risultati di una query sulle collezioni di documenti. Ogni documento che risulta rilevante per una ricerca viene, infatti, presentato all'utente assieme ad alcune delle

informazioni strutturate disponibili relative ad esso ed al progetto al quale appartiene.

L'interfaccia della ricerca è stata progettata in modo tale da valorizzare i dati strutturati forniti. I dati più importanti riguardanti ciascun documento sono stati posti in risalto, rendendoli sempre visibili; ulteriori informazioni sono state poi rese visualizzabili per ogni risultato dando all'utente la possibilità di farle comparire nel caso in cui volesse avere maggiori dettagli su uno o più risultati specifici della ricerca.

Si è scelto di mostrare, per ogni documento nella lista dei risultati, insieme al suo titolo, le seguenti informazioni: una breve descrizione del documento, il progetto al quale esso appartiene e il budget complessivo assegnato al progetto (Figura 4.1).

Figura 4.1 Un esempio di risultato di ricerca nel quale sono visibili i principali metadati relativi al documento.

Le informazioni aggiuntive visualizzabili su richiesta dell'utente per ogni documento sono state suddivise in informazioni relative al deliverable e informazioni relative ai partner del progetto nell'ambito del quale il documento è stato scritto; i due sottogruppi di informazioni si possono visualizzare separatamente. I dati relativi al documento comprendono: la data in cui esso è stato scritto, il budget assegnato ad esso nel contesto del progetto cui afferisce, le parole chiave utilizzate per descrivere il documento e l'elenco delle categorie di soggetti istituzionali, pubblici e privati, ai quali il documento è indirizzato. I dati relativi ai partner di progetto sono, invece: il nome del partner, che viene indicato come "lead

partner" quando è il partner capofila del progetto, il paese di provenienza del partner e il budget assegnato ad esso nell'ambito del progetto (Figura 4.2).

Figura 4.2 Un esempio di risultato di ricerca nel quale sono visibili tutti i metadati relativi al documento.

L'integrazione di informazioni strutturate nei risultati della ricerca semantica abilita l'utente ad un'analisi più semplice ed efficace dei risultati di una ricerca. In un contesto reale di utilizzo del sistema, infatti, anche un utente esperto non avrebbe modo di ricordare il contesto di provenienza di tutti i possibili documenti che potrebbe ricevere come risultati di una ricerca. I metadati che sono mostrati insieme ad essi danno, invece, all'utente la possibilità di inquadrare il contesto nel quale ciascun documento si inserisce, avendo quindi un'idea più precisa, ad esempio, del tema trattato o tenendo presente il progetto nell'ambito del quale è stato prodotto. Questo consente all'utente sia di valutare meglio se un documento sia o meno di proprio interesse rispetto alla ricerca compiuta, che di tener presente altri aspetti legati alla propria ricerca, osservando, ad esempio, quali siano i principali progetti e partner impegnati sul tema da lui ricercato, o quale sia la portata degli investimenti stanziati su di esso.

4.2 Informazioni strutturate per la matrice di correlazione tra concetti

La matrice di correlazione tra concetti di una collezione, descritta nel paragrafo 2.3.2 è un servizio basato esclusivamente sulla rappresentazione semantica dei documenti e dei concetti contenuti in essi. Data una coppia di concetti – o più coppie, a seconda delle dimensioni scelte per la matrice – il sistema compie una ricerca nello spazio multidimensionale dei vettori che rappresentano i termini, secondo le tecniche di rappresentazione e di confronto tra termini descritte nei paragrafi 1.3.1 e 1.3.2. In particolare, il sistema ricerca i concetti più simili a ciascuno dei concetti della coppia fornita; le due liste di concetti risultanti dalle ricerche vengono poi unite utilizzando l'algoritmo combSUM. I primi dieci risultati più rilevanti vengono poi restituiti all'utente riempiendo le celle della matrice corrispondenti ad ogni coppia di concetti.

Ogni concetto risultante dalla ricerca e inserito nella matrice è stato progettato come un link che porti l'utente direttamente ad una ricerca semantica di quel concetto nella collezione sulla quale la matrice è stata costruita. In questo modo, l'utente può esplorare ulteriormente il contenuto della collezione visualizzando i documenti che più sono attinenti ad un concetto di interesse all'interno di essa.

Grazie al collegamento stabilito tra il servizio di correlazione tra concetti in una collezione e la ricerca semantica all'interno di essa, i metadati sul dominio possono quindi ampliare anche le possibilità offerte dal servizio di correlazione tra concetti.

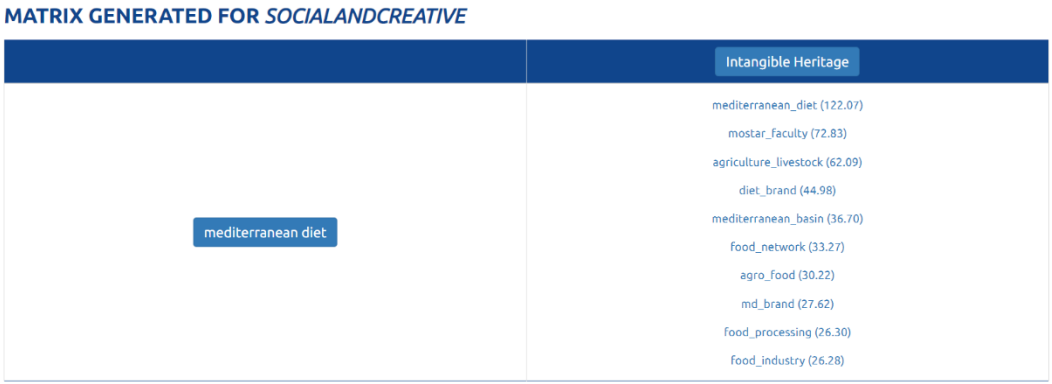

Figura 4.3 Un esempio di matrice 1x1 generata sulla collezione della Community Social and Creative

4.3 Informazioni strutturate per la creazione di mappe tematiche

Come accennato all'inizio del capitolo, il sistema di supporto alle decisioni sviluppato comprende un servizio di visualizzazione geografica dei risultati di una ricerca semantica. Il suo fine è facilitare l'esplorazione delle relazioni possibili o esistenti tra i progretti o i partner grazie ad una visualizzazione che unisce le potenzialità della rappresentazione semantica del testo, la precisione delle informazioni strutturate disponibili sul dominio e l'immediatezza della rappresentazione geografica dei risultati.

L'idea alla base del servizio è quella di facilitare l'interpretazione dei risultati di una ricerca, partendo dall'esplicitare la collocazione geografica dei partner che sono coinvolti nei progetti risultanti o che sono autori dei documenti stessi. La scelta di adottare la visualizzazione geografica come componente principale di questo servizio è dovuta alla caratteristica

propria dei progetti del programma Med di essere distribuiti su tutta l'area del Mediterraneo. Individuare precisamente i luoghi in cui operano i partner può consentire ai policy maker di avere una maggiore consapevolezza di quali siano i territori coinvolti negli investimenti, e ai partner di progetto di individuare potenziali soggetti con cui collaborare in futuro in virtù dell'eventuale vicinanza geografica.

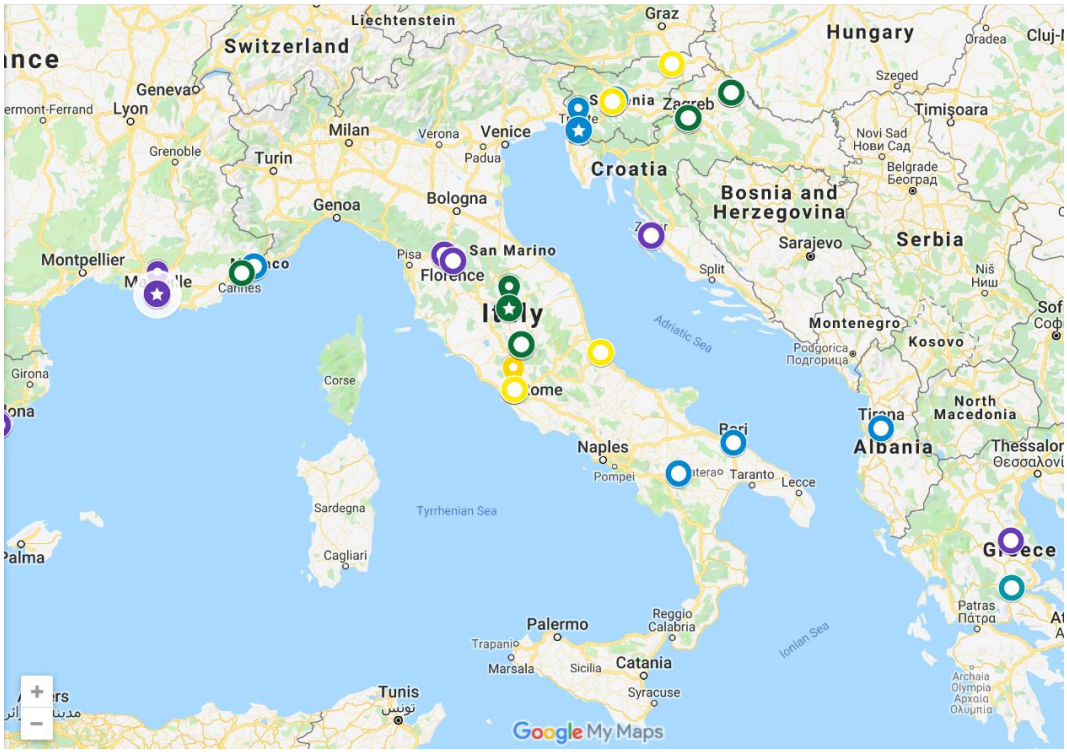

Figura 4.4 Una mappa che mostra la collocazione dei partner di diversi progetti (mostrati con colori diversi)

L'accesso flessibile alle informazioni strutturate è un prerequisito essenziale per la generazione dinamica e automatica delle mappe associate ai risultati di ricerca. In primo luogo, i dati relativi alla collocazione geografica dei partner sono ovviamente fondamentali per la creazione delle mappe, e sono facilmente ottenibili grazie alle API REST di gestione dei metadati esposte dal Semantic Framework. Allo stesso modo si può ottenere l'informazione che specifica il progetto cui un partner sta partecipando, codificata nella mappa usando indicatori di colori diversi per ogni progetto.

Il servizio di visualizzazione geografica dei risultati non si limita però alla visualizzazione della collocazione geografica dei partner e dei progetti relativi ai documenti risultanti da una ricerca. Grazie all'accesso flessibile alle informazioni strutturate, garantito dai servizi di gestione dei metadati integrati nel Semantic Framework, è possibile infatti visualizzare sulla mappa moltissime informazioni diverse, relative tanto ai singoli documenti, quanto ai progetti in cui sono stati scritti e ai partner coinvolti in essi.

Sono stati, dunque, progettati tre diversi tipi di mappe che è possibile visualizzare a partire dai risultati di una ricerca semantica:

- Una mappa che mostra i partner di tutti i progetti nell'ambito dei quali sono stati prodotti i documenti risultanti dalla ricerca, ossia i partner che lavorano sul tema cercato; in questa mappa è possibile anche evidenziare i partner a capo dei singoli progetti e quelli che sono autori dei documenti presenti nel ranking.

- Una mappa che mostra il budget che ciascuno di questi partner sta investendo sul tema ricercato;

- Una mappa che mostra la posizione geografica degli stakeholder coinvolti nei progetti cui appartengono i documenti che sono rientrati tra i risultati della ricerca.

Per posizionare sulla mappa gli elementi che si vogliono visualizzare (partner, budget investito o stakeholder) il servizio esterno di creazione delle mappe può avvalersi dei dati strutturati relativi a ciascuna di queste entità.

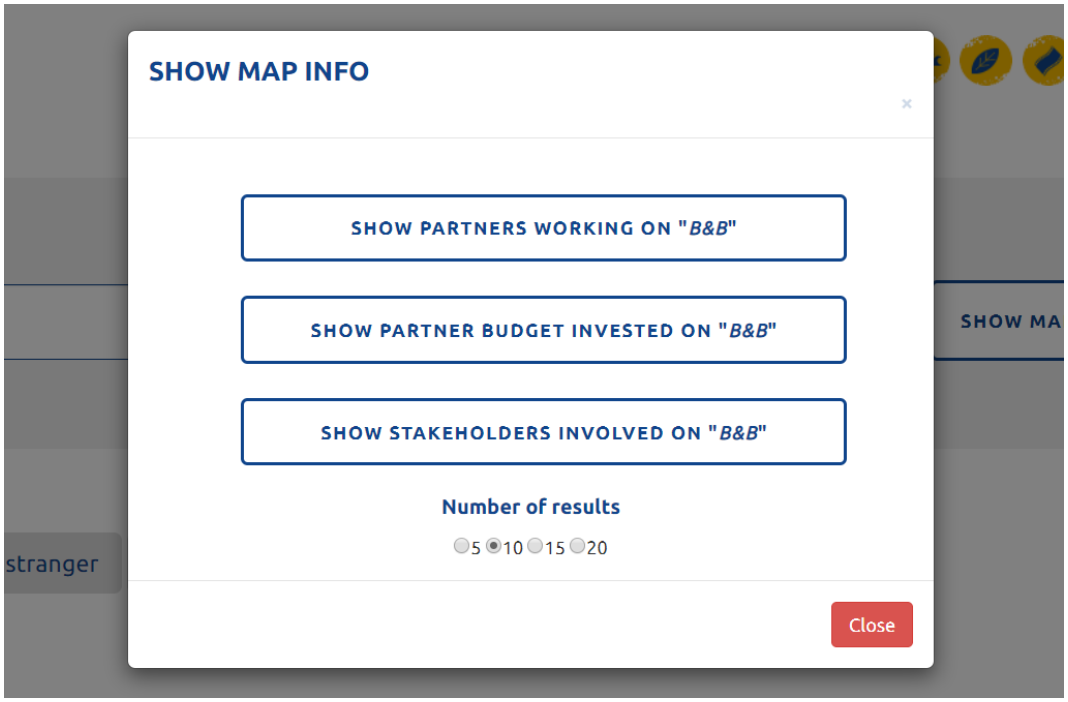

Figura 4.5 Pannello di scelta del tipo di mappa da visualizzare sulla base dei risultati della ricerca "B&B"

La visualizzazione geografica può impiegare i dati strutturati arricchendo la visualizzazione geografica in diversi modi. Nella prima tipologia di mappe, ad esempio, può essere integrata nella visualizzazione l'informazione riguardo le Community cui afferisce ogni progetto mostrato; o, ancora, si può facilmente mostrare graficamente l'informazione riguardo l'arco temporale di svolgimento di un progetto, indicando quindi se questo sta volgendo al termine o meno.

Nel caso del secondo tipo di mappe descritto, l'informazione stessa sul budget investito sul tema cercato, che corrisponde al budget assegnato al deliverable che è risultato nella ricerca, può essere messa in relazione con il budget totale assegnato, nell'ambito del progetto, al partner che ha scritto il documento, nonché con il budget complessivo assegnato al progetto (entrambe informazioni disponibili nella base di dati). In questo modo, sarebbe possibile compiere una valutazione comparativa della spesa dei

diversi partner e nei diversi progetti su un tema, valutando quindi l'andamento dell'investimento compiuto.

Tanto nel primo tipo di mappa, quanto nel terzo, inoltre, può risultare utile visualizzare graficamente la tipologia di partner e stakeholder impegnati o coinvolti in un progetto (campi *partner_nature* e *stakeholder_nature* nelle rispettive tabelle del database). In questo modo è possibile, nel primo caso, individuare la tipologia di soggetti impegnati nei progetti che operano sul tema, e quale sia, ad esempio, la categoria di soggetti che ha scritto più documenti relativi ad un tema ricercato. Nel secondo caso, ossia nella mappa di visualizzazione degli stakeholder, può risultare utile comprendere quali siano le categorie di soggetti maggiormente coinvolte dai progetti che operano su un certo tema, anche in rapporto, ad esempio, alla collocazione geografica dei partner che vi lavorano.

Grazie all'interfaccia progettata, mostrata in Figura 4.5, l'utente può anche selezionare il numero di risultati di ricerca che desidera visualizzare sulla mappa creata, avendo così la possibilità di creare mappe più o meno complesse e ricche di informazioni.

Grazie al servizio di visualizzazione dei risultati di ricerca l'utente ha quindi diversi modi per esplorare la realtà dei progetti nel programma Med in relazione ad un particolare tema di suo interesse, analizzando in modo dettagliato informazioni su chi stia operando su quel tema e dove, quale sia l'entità degli investimenti stanziati e quali siano i soggetti coinvolti nelle azioni intraprese dai diversi progetti.

5. Conclusioni e svluppi futuri

Il progetto TALIA ha avuto come obiettivo il miglioramento del coordinamento e della comunicazione tra soggetti operanti nell'area del Mediterraneo attraverso lo sviluppo del sistema di supporto alle decisioni illustrato nel presente lavoro di tesi. In particolare, il sistema di supporto alle decisioni, basato sul Semantic Framework, offre all'utente la possibilità di acquisire conoscenza nel dominio dei progetti del programma Med attraverso l'utilizzo del motore di ricerca semantico, del servizio di analisi delle correlazioni esistenti tra i concetti di una collezione, e delle informazioni strutturate sul dominio integrate nei servizi del sistema come illustrato nel capitolo precedente.

Lo sviluppo della funzionalità di ricerca semantica come servizio web, nonché la cura dell'interfaccia utente del servizio, ha assunto un ruolo importante nel rendere tale servizio accessibile agli utenti in modo più semplice. Le funzionalità di ritrovamento di concetti simili a quelli presenti nella query dell'utente e la possibilità di esplorare le collezioni attraverso le ricerche svolte a partire dai concetti ritrovati, può supportare notevolmente l'utente nel soddisfare il bisogno informativo espresso tramite una query esplorando le connessioni semantiche presenti nelle collezioni. La rappresentazione semantica dei testi è stata, inoltre, utilmente sfruttata nell'implementazione del servizio di "Summarization", che permette all'utente di visionare una parte ridotta del contenuto di ogni documento.

Il caso d'uso della matrice di correlazione tra concetti dà all'utente la possibilità di esplorare in maniera ancora più ampia il contenuto informativo presente nelle collezioni considerate, mettendo in evidenza, a prescindere dai singoli documenti presenti in esse, quali siano le connessioni tra i temi trattati in una intera collezione.

Infine, l'integrazione di dati non strutturati relativi al dominio delle collezioni, organizzati in una base di dati relazionale, ha ulteriormente arricchito la disponibilità di informazioni a disposizione dell'utente del sistema, pur mantenendo flessibile rispetto alle sue esigenze la quantità e il tipo di informazioni di volta in volta mostrate, grazie alla progettazione di una interfaccia utente attenta alle esigenze di usabilità del software. L'integrazione di un modulo di gestione dei dati strutturati ha permesso, inoltre, l'implementazione del caso d'uso di visualizzazione geografica dei risultati di ricerca, consentendo un accesso ai metadati più flessibile rispetto alla tecnica inizialmente adottata, che consisteva nel memorizzarli assieme al corpo del documento come parti di esso, secondo la strutturazione dei documenti disponibile nella libreria Lucene.

Gli sviluppi futuri del progetto di tesi sono molteplici, sia dal punto di vista del miglioramento tecnico delle funzionalità del Semantic Framework, che da quello di possibili ulteriori scenari nei quali il sistema potrebbe essere impiegato.

Occorre, innanzitutto, evidenziare che tanto la piattaforma "Semantic Framework", quanto i servizi del sistema di supporto alle decisioni, offrono funzionalità declinabili alle esigenze di una qualsiasi organizzazione e a qualsiasi dominio applicativo. Le funzionalità di gestione di collezioni di documenti e di analisi semantica del testo offerte dal Semantic Framework sono adatte all'implementazione di molti servizi che necessitino di un'analisi di testi approfondita. Analogamente, come è stato illustrato inizialmente, la ricerca semantica e la matrice di correlazione tra concetti sono servizi innovativi che possono essere molto significativi in un sistema di supporto alle decisioni basato su dati testuali sviluppato per un qualsiasi dominio.

Nonostante il progetto TALIA avesse come obiettivo quello di favorire lo sviluppo della comunità Social and Creative, è evidente – anche dalle considerazioni appena fatte – che le analisi abilitate dal sistema di supporto alle decisioni possano essere facilmente applicate alle collezioni di documenti attinenti alle altre Community, in modo tale da supportarne la crescita.

È opportuno, tuttavia, che in futuro il sistema sia integrato in modo più efficiente con le fonti documentali e informative relative al dominio dei progetti del programma Med. Oltre a semplificare il processo di ritrovamento e aggiunta dei documenti alle collezioni, e di inserimento dei metadati nella base di dati, il sistema potrebbe anche beneficiare della presenza di una maggiore quantità di documenti, che renderebbero le operazioni di analisi dei documenti ancor più significative.

Dal punto di vista dei servizi esposti dal sistema di supporto alle decisioni, è possibile implementare alcuni miglioramenti tecnici. Nel motore di ricerca semantico, ad esempio, può essere integrata una funzione di relevance feedback che permetta all'utente di notificare al sistema quando un documento non è rilevante rispetto alla query che ha svolto; ciò porterebbe al raffinamento della funzione di scoring del motore di ricerca e, dunque, ad ottenere dei risultati di ricerca maggiormente rilevanti.

I risultati attualmente ottenuti tramite il servizio di correlazione tra concetti potrebbero essere confrontati con quelli ottenuti implementando metodi diversi da quello utilizzato per combinare le liste di concetti simili ritrovati dai metodi implementati all'interno del Semantic Framework. L'efficacia di questo caso d'uso può essere, inoltre, misurata effettuando sperimentazioni che coinvolgano esperti del dominio chiamati ad interpretare i risultati

ottenuti dalla correlazione dei concetti definiti, così da poter identificare, ed eventualmente correggere, la presenza di risultati indesiderati.

Infine, il modulo di gestione dei metadati potrebbe essere reso indipendente dal dominio, in modo tale da poter gestire le informazioni strutturate attinenti alle collezioni presenti nel Semantic Framework indipendentemente dalla struttura della base di dati nella quale sono memorizzate.

Riferimenti bibliografici

[1] Porter, M. F., "An algorithm for suffix stripping", Program, Vol. 14 Issue: 3, pp.130-137, 1980

[2] Landauer, T. K. and Dumais, S. T., "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." , Psychological review, vol. 104, no. 2, pp. 211, 1997

[3] Lancaster, F. W. and Fayen, E. G., "Information Retrieval On-Line", Los Angeles: Wiley-Becker & Hayes, 1974

[4] Baeza-Yates, R., Ribeiro-Neto, B., "Modern Information Retrieval", Harlow, England: Pearson Addison Wesley. ISBN: 978-0-321-41691-9, 2011

[5] Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", IBM Journal of Research and Development. 1 (4): 309–317, 1957

[6] Spärck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation. 28: 11–21, 1972

[7] Ramos, J. et al., "Using tf-idf to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, 2003

[8] Salton, G., Wong A., Yang C. S., "A vector space model for automatic indexing", in Communications of the ACM, Volume 18 Issue 11, pp. 613-620, Nov. 1975

[9] Harris, Z., "Distributional structure", Word. 10 (23), pp. 146–162, 1954

[10] Johnson, W. and Lindenstrauss, J., "Extensions of Lipschitz mappings into a Hilbert space, in Contemporary Mathematics", American Mathematical Society, vol. 26, pp. 189–206, 1984

[11] Sahlgren M., "An Introduction to Random Indexing," in Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE, vol. 5, 2005

[12] Fielding R. T., "Architectural Styles and the Design of Network-based Software Architectures", 2000

[13] Massè, M., "REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces", O'Reilly, 2012

[14] Rossiello G., Basile P., e Semeraro G., "Centroid-based text summarization through compositionality of word embeddings", In Proc. of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pages 12–21, 2017

Ringraziamenti

Il primo grazie va alla mia famiglia, senza la quale non sarei arrivato fin qui.

Vi ringrazio per avermi cresciuto stimolando sempre la mia curiosità, la forza che mi ha permesso di non fermare mai il mio cammino in questo percorso, grazie alla quale ho trovato di ogni cosa la prospettiva più interessante.

A mio padre, per avermi insegnato, con l'esempio, la gentilezza e la dedizione, senza le quali non sarei chi sono.

A mia madre, che mi ha insegnato l'indipendenza e il saper contare su sé stessi, le fondamenta sulle quali ho costruito e intendo costruire.

A mia sorella, per il suo essere "altro da me" e per aver portato, da sempre, energia e gioia a movimentare la calma dei miei pensieri.

A mio nonno Salvatore, per avermi accompagnato, pranzo dopo pranzo, fino a questo giorno. Grazie per i tuoi racconti e per il tuo esempio di vita.

Il secondo grazie va a Fabiola. Senza di te, come sai, questo percorso non sarebbe mai stato lo stesso. Grazie di essermi sempre stata vicino, di avermi supportato e aver condiviso i miei tanti dubbi e le mie gioie, di avermi criticato ogni volta che era necessario, di aver sempre voglia di darmi fastidio e prendermi in giro. Grazie di aver fiducia in me. Come sai, sei la migliore.

Voglio ringraziare il professor Lops, in primo luogo per la dedizione e la passione dimostrate nella sua attività di docente, qualità non comuni, che sono state per me una luce nel percorso di studi. Lo ringrazio, in secondo luogo, per averci dato l'opportunità di svolgere questo lavoro di tesi, di grande valore formativo, che ha aperto, per me, porte su mondi nuovi.

Ringrazio il dottor Pierpaolo Basile, per la sua gentilezza, disponibilità ed ironia, che ci hanno accompagnato negli ultimi mesi.

A Studenti Indipendenti va un ringraziamento speciale. Grazie a voi sono certo avrò sempre un posto da poter chiamare casa all'Uniba, in qualunque plesso mi trovi. Anche a Economia. Vi ringrazio per avermi dato fiducia fin dall'inizio. Grazie per la vostra energia e per la vostra capacità di confronto aperto e franco. Che ciò possa essere sempre fonte di cambiamento, di coinvolgimento e di attivazione degli studenti della nostra Università.

Ai ragazzi di Informatica: la nostra piccola famiglia è appena nata, davanti a voi c'è la grande sfida e opportunità di farla crescere. Trovate ciò che vi motiva a migliorare, ciò che vi stimola, ciò che vi appassiona, seguitelo e poi condividetelo con gli altri. Ricordate sempre di ascoltarvi a vicenda, di mettere voi stessi in discussione, e di farlo con gli altri in modo costruttivo; ricordate di aver cura delle persone che condividono questa esperienza bellissima con voi, cercando di capire chi vi sta di fronte prima di giudicare. Ricordatevi, ogni tanto, anche di studiare.

Allo staff del BarCollo, grazie per tutto quello che abbiamo condiviso fin qui. Cosa sarebbero stati questi tre anni senza le entusiasmanti uscite a Storie del sabato sera? Ma anche del venerdì sera, e a volte anche della domenica o del giovedì… per fortuna ci siete sempre stati voi.

A Fernando e Lorenzo, per il supporto reciproco e l'impegno che ci ha unito in questi mesi, e per l'apertura e la disponibilità che avete mostrato nei miei confronti.