Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il Management

Un ambiente estensibile per annotare citazioni e riferimenti bibliografici

Relatore:
Prof. Angelo Di Iorio

Presentata da:
Francesco Guerrini

Sessione III
Anno Accademico 2021-2022

Indice

Elenco delle figure

Introduzione

Nel web esistono enormi raccolte di articoli scientifici resi pubblici e raggiungibili da tutta la comunità. Per raccogliere informazioni per una nuova ricerca è solito consultare l'elaborato di altri ricercatori, per farlo si utilizzano diversi tool come portali web o motori di ricerca come Google Scholar. Per migliorare una ricerca esistono ulteriori strumenti e indicazioni per trovare articoli e autori affidabili: vengono spesso consultati degli indicatori bibliometrici, ovvero misurazioni relative alle pubblicazioni scientifiche e al loro impatto all'interno della comunità scientifica.

La citazione costituisce un legame tra due articoli: rappresenta l'idea o il pensiero sviluppato da un autore riportato nel testo di un altro autore. Esse sono utilizzate per fornire supporto ad una propria affermazione, per attribuire il merito per un'idea originale ad un'altra persona, o per mostrare la fonte di un'informazione. Inoltre, le citazioni possono anche essere utilizzate per mostrare che l'autore del testo sta considerando le idee degli altri e le sta utilizzando nella propria argomentazione.

L'importanza di un articolo di ricerca viene regolarmente misurata contando quante volte è stato citato. Tuttavia, trattare tutte le citazioni con lo stesso peso ignora l'ampia varietà di funzioni che svolgono.

È importante conoscere i diversi motivi per cui un articolo viene citato perché ciò ci permette di comprendere meglio le dinamiche della comunicazione scientifica e di valutare l'impatto di un'opera su un dato campo di ricerca. Inoltre, la conoscenza delle funzioni citazionali può essere utilizzata per migliorare l'organizzazione e la disseminazione della letteratura scientifica, ad esempio attraverso la creazione di sistemi di indicizzazione e di ricerca più efficaci e personalizzati per gli utenti.

Per questo emerge la necessità di analizzare le funzioni citazionali, esaminando il contesto interno delle citazioni e cercando di estrapolarne un intento; a questo fine sono stati realizzati diversi modelli di machine learning per la classificazione delle funzioni citazionali.

La ricerca sull'analisi della funzioni citazionali è in continuo sviluppo, sono stati fatti diversi studi e proposti diversi modelli per l'apprendimento automatico e la classificazione di tali funzioni.
Questo ambito di ricerca però è molto eterogeneo, infatti ogni studio utilizza mezzi diversi e porta anche a risultati differenti.
La differenza tra i risultati è causata principalmente dalla natura delle citazioni e dalla qualità e quantità dei dati utilizzati per addestrare i modelli.

Questa tesi nasce dalla necessità di dare un contributo alla ricerca, tenendo conto dell'eterogeneità dell'ambito di studio, fornendo un ambiente per l'annotazione manuale che permetta di raccogliere dei dati validi per lo sviluppo della annotazione automatica.

Capitolo 1

Analisi delle funzioni citazionali: stato dell'arte

1.1 Introduzione alla ricerca sulle funzioni citazionali

L'analisi citazionale è un metodo di analisi quantitativa che utilizza i dati delle citazioni per valutare l'impatto e l'influenza delle pubblicazioni scientifiche e degli autori. Utilizzando strumenti e database specifici, gli studiosi possono raccogliere e analizzare i dati delle citazioni per generare indicatori di impatto come l'H-index e il G-index. Le relazioni tra le pubblicazioni e tra gli autori e le tendenze nei vari ambiti vengono studiati attraverso l'analisi delle co-citazioni (due o più risorse citate insieme da una stessa fonte) e dalle reti di citazioni. Le misure bibliometriche danno una valutazione puramente quantitativa, contando quanti articoli citano un dato articolo [1] o mediante misure più sofisticate come l'H-index [2].
Tuttavia è riconosciuto che le citazioni hanno diversi scopi. I ricercatori hanno da tempo affermato che la classificazione delle citazioni è un elemento centrale per comprendere la rilevanza dell'articolo nel campo quindi è necessario tenere conto del contenuto e del contesto delle citazioni. Bonzi [3], fa notare che le citazioni non sempre significano che il lavoro citato sia valutato positivamente, e Ziman [4] afferma che molte citazioni sono fatte per "cortesia" (verso potenti approcci rivali), "politica" (per nome e argomento per autorità) o "pietà" (verso i propri amici, collaboratori e superiori). I ricercatori spesso seguono anche l'usanza di citare alcuni articoli che semplicemente forniscono le basi del loro argomento attuale ("rendere omaggio ai pionieri").
Lo studio delle funzioni citazionali è quindi una branca dell'analisi citazionale che si concentra sull'identificazione dei motivi per cui un autore cita un particolare documento in un testo.
Ci sono diverse etichette che vengono utilizzate per descrivere le funzioni citazionali, tra cui "supporto", "contestualizzazione", "riproposizione" e "critica". Ad esempio, una citazione può essere utilizzata per fornire supporto a un'affermazione o per contestualizzare un'idea all'interno del lavoro più ampio. In altri casi, una citazione può essere utilizzata per riproporre un'idea già presente in un lavoro precedente o per criticare un'affermazione o un'idea.
Per identificare le funzioni citazionali, vengono utilizzate diverse tecniche di elabo-

razione del linguaggio naturale e di apprendimento automatico, come avviene nel modello ACL-ARC. Lo studio delle funzioni citazionali è importante perché può fornire informazioni su come un documento è stato utilizzato e influenzato dai lavori che lo hanno preceduto e possono essere utilizzate per comprendere meglio il contesto e l'impatto della ricerca accademica.

1.2 Storia dell'analisi citazionale

La storia dell'analisi citazionale può essere tracciata fino agli anni '20 e '30 del secolo scorso, con l'emergere degli studi bibliometrici. Gli studi bibliometrici erano principalmente concentrati sulla quantificazione delle pubblicazioni scientifiche e delle citazioni per valutare l'impatto del lavoro scientifico.

Uno dei precursori della ricerca bibliometrica è stato il matematico e statistico belga Paul Otlet, che ha sviluppato un sistema di classificazione chiamato "Universal Decimal Classification" (UDC) per organizzare le informazioni scientifiche. Nel suo libro "Monde: Essai d'universalisme" [5], Otlet ha proposto un sistema globale di comunicazione e documentazione basato sulla UDC, che avrebbe permesso di accedere alle informazioni scientifiche in modo efficiente.
Un altro precursore degli studi bibliometrici è stato il bibliotecario e bibliometrista americano Eugene Garfield, che ha fondato la rivista "Science Citation Index" (SCI) nel 1961. La SCI ha fornito un metodo per individuare le citazioni tra le pubblicazioni scientifiche, permettendo agli scienziati di determinare l'impatto delle loro pubblicazioni e di identificare i lavori più influenti nel loro campo.
Gli studi bibliometrici hanno dato origine ad indicatori di impatto come l'indice di Hirsch (H-index), che misura la produttività e l'impatto di un autore in base al numero di pubblicazioni e di citazioni ricevute. L'indice di Hirsch è stato proposto per la prima volta nel 2005 dallo scienziato J.E. Hirsch e si è dimostrato uno strumento efficace per valutare l'impatto del lavoro scientifico degli autori.

Nel corso degli anni, la ricerca citazionale si è evoluta per includere l'uso di tecnologie e metodi analitici avanzati, come l'analisi delle reti di co-citazione, l'analisi delle tendenze, l'analisi semantica e l'apprendimento automatico, per analizzare i dati delle citazioni e comprendere come le idee si diffondono all'interno della comunità scientifica.
Inoltre, con l'aumento della disponibilità dei dati delle citazioni, gli scienziati sono stati in grado di utilizzare l'analisi citazionale per identificare i leader e gli innovatori nel loro campo, mappare le relazioni tra le pubblicazioni e gli autori, e comprendere meglio il processo di diffusione delle conoscenze.
Alla fine degli anni '90 e all'inizio del 2000, la ricerca citazionale si è evoluta ulteriormente con l'emergere dell'analisi delle funzioni citazionali. Questa area di ricerca si concentra ora sullo studio del significato e dello scopo delle citazioni all'interno dei testi scientifici. Gli scienziati hanno iniziato a utilizzare tecniche qualitative e quantitative per analizzare il contesto e le intenzioni dietro alle citazioni, piuttosto che semplicemente contarne il numero.

In sintesi, la storia dell'analisi citazionale si è evoluta attraverso una serie di fa-

si, dalla quantificazione delle pubblicazioni scientifiche e delle citazioni negli studi bibliometrici, alla valutazione dell'impatto del lavoro scientifico attraverso indicatori di impatto, all'uso di tecnologie e metodi analitici avanzati per comprendere il processo di diffusione delle conoscenze, fino all'analisi delle funzioni citazionali per comprendere il significato e lo scopo delle citazioni all'interno dei testi scientifici.

1.3 Le applicazioni delle funzioni citazionali

È importante conoscere i diversi motivi per cui un articolo viene citato perché ciò può fornire un'indicazione sulla qualità e l'impatto di quel lavoro. Ad esempio, le citazioni che utilizzano un lavoro come riferimento o come fonte di informazioni su un particolare argomento possono indicare che il lavoro è considerato una fonte autorevole di informazioni. D'altra parte, le citazioni che criticano o contestano un lavoro possono indicare che quel lavoro ha limitazioni o che ci sono opinioni divergenti in merito all'argomento.
Inoltre, comprendere le funzioni delle citazioni può anche essere utile per gli autori, in quanto può aiutare a migliorare la qualità del proprio lavoro. Ad esempio, conoscere le ragioni per cui un lavoro viene citato può aiutare gli autori a scegliere le fonti adeguate per sostenere le loro argomentazioni o a individuare le aree in cui il loro lavoro potrebbe essere migliorato.
La ricerca sulle funzioni citazionali può fornire importanti informazioni sulle interazioni tra le idee all'interno di un campo di ricerca e possono contribuire in vari modi:

- **Costruzione di reti citazionali qualitative**
  L'utilizzo delle funzioni citazionali migliora notevolmente i collegamenti tra due articoli specificando il motivo per cui essi sono relazionati. La rete citazionale può mettere in evidenza il percorso di un autore valorizzandolo; per esempio, mostrando quali articoli ha pubblicato, da quali ha preso spunto, quelli che hanno preso spunto da lui e per quale motivo. In questo modo si riesce a valorizzare il lavoro di un autore e riconoscere il tipo di contributo che sta dando in un determinato ambito. Non più solo tramite il numero di citazioni collegate a lui ma con il tipo di contributo che dà alla ricerca.

- **Identificazione delle fonti autorevoli**
  Le funzioni citazionali possono essere utilizzate per identificare le fonti più autorevoli all'interno dell'area di ricerca quando le citazioni vengono utilizzate a supporto di un'affermazione e/o per contestualizzare, piuttosto che come critica e confutazione.

- **Indicizzazione più efficace degli articoli**
  La conoscenza delle funzioni citazionali può essere utilizzata per migliorare l'organizzazione e la disseminazione della letteratura scientifica, ad esempio attraverso la creazione di sistemi di indicizzazione e di ricerca più efficaci e personalizzati per gli utenti. Per esempio, se si vogliono raccogliere dei dati iniziali per un nuovo studio, conoscendo la funzione che può avere una citazione, si potrebbero estrarre direttamente solo gli articoli che propongono lo studio e l'elaborazione di dati.

Un esempio di applicazione attuale è Semantic Scholar[1], un motore di ricerca accademico sviluppato da Allen Institute for Artificial Intelligence (AI2). Tramite Semantic scholar è possibile ricercare milioni di documenti scientifici e su ognuno di questi visualizzare delle informazioni come ad esempio l'abstract, le citazioni, le referenze bibliografiche e le informazioni sull'autore. Questo motore di ricerca, oltre a identificare il numero di citazioni ricevute per un articolo, mostra anche la funzione che queste esprimono. Allo stesso modo mostra il motivo per cui l'autore dell'articolo menziona le proprie fonti.
Nella figura 1.1 mostriamo un esempio di articolo ricercato tramite Semantic Scholar. Possiamo notare che a destra vengono riportate tutte le citazioni su questo articolo (136) e sottostante le funzioni che esprimono, queste sono divise in tre categorie: "Background","Methods","Results". Nell'anteprima dell'articolo, nella parte di testo evidenziata, vediamo che l'autore cita un'altra fonte e il motore di ricerca ci suggerisce che lo scopo di tale riferimento è "Methods".

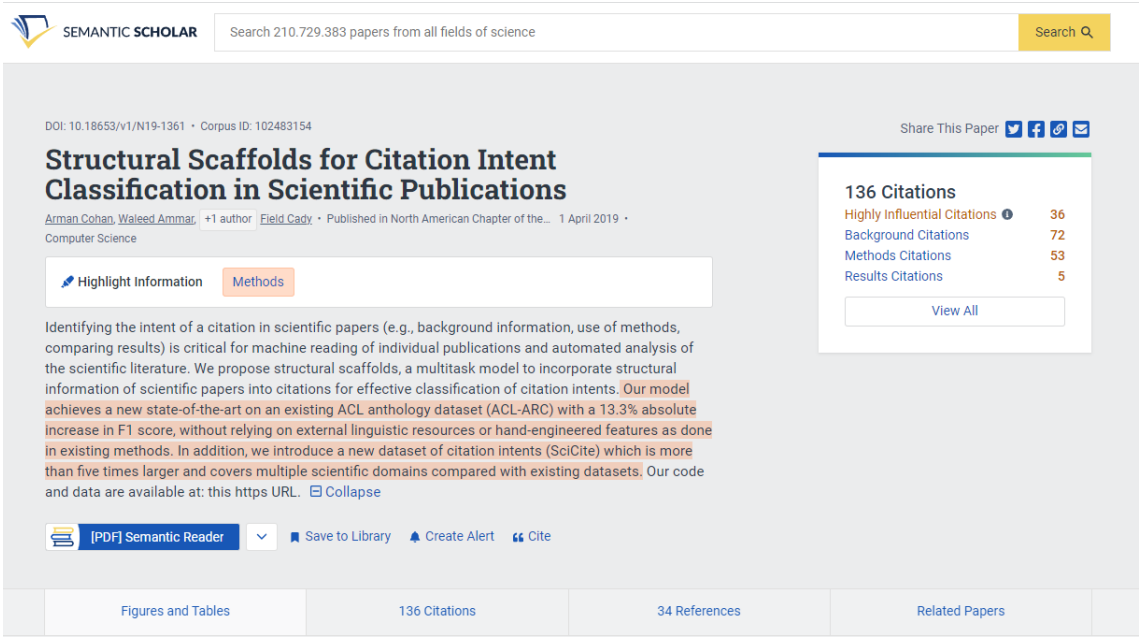


Figura 1.1: Le funzioni citazionali di Semantic Scholar

1.4 I classificatori di funzioni citazionali

La ricerca sulle funzioni citazionali è in continua evoluzione e nel tempo si stanno riscontrando diversi studi e diverse possibili soluzioni. Nello specifico sono stati proposti più modelli di machine learning e classificatori che sono in grado di identificare e categorizzare automaticamente le citazioni in un testo in base alle funzioni che svolgono. I modelli proposti spesso utilizzano un algoritmo di apprendimento automatico, come ad esempio una rete neurale artificiale o un albero di decisione, per identificare pattern e schemi nei dati di addestramento e associare questi pattern

[1]Semantic Scholar: https://www.semanticscholar.org/

alle funzioni citazionali specifiche.
I primi lavori sulla classificazione automatizzata dell'intento di una citazione erano realizzati su sistemi basati su regole, come ad esempio quelli proposti da Garzone e Mercer [6] e Pham e Hoffmann [7]. Successivamente i metodi di apprendimento automatico basati su modelli linguistici e altre funzionalità che considerano il contesto delle citazioni si sono rivelati efficaci per esempio come l'uso di "frasi chiave" proposto da Tefel [8]. Il modello di Abu-Jbara [9] si basava su caratteristiche lessicali, strutturali e sintattiche e su un SVM lineare per la classificazione.

Ci troviamo davanti ad un ambito di ricerca molto eterogeneo, esistono infatti diversi classificatori progettati per diversi casi d'uso. Vediamo ora due esempi di classificatori sviluppati recentemente: ACL-ARC e SciCite.

ACL-ARC

Il modello ACL-ARC [10] è una nuova proposta per la classificazione della funzione di una citazione che utilizza una combinazione di embedding di parole contestuali basati sull'autore e sul titolo degli articoli, così come una rete neurale per classificare il motivo di una citazione. Gli embedding di parole contestuali sono una rappresentazione matematica del significato di una parola all'interno di un contesto specifico. Consistono nel mappare ogni parola in un vettore di numeri reali in modo che le parole simili siano rappresentate da vettori simili.
La combinazione di embedding di parole contestuali di autore e di titolo riesce a rappresentare il contesto più ampio dell'articolo citato, questo permette al modello di prendere in considerazione il contesto delle parole non solo in base alla loro posizione nel testo, ma anche in base al contesto generale dell'articolo. Il modello è stato addestrato su un insieme di dati di citazioni mostrando di migliorare significativamente l'accuratezza della classificazione rispetto ai modelli esistenti.
Il processo completo del modello ACL-ARC per estrarre una funzione citazionale consiste in diverse fasi. In primis, i dati di input vengono pre-elaborati per estrarre le citazioni e i metadati degli articoli. Successivamente, gli embedding di parole contestuali vengono creati utilizzando una tecnica di apprendimento non supervisionato su grandi quantità di testo.
In seguito il modello viene addestrato su un insieme di dati di citazioni, e una volta addestrato, può essere utilizzato per classificare il motivo di una citazione in un testo.

La classificazione ACL-ARC cattura le ampie funzioni tematiche che una citazione può svolgere in un discorso e le differenzia in sei etichette.

Class	Description	Example
BACKGROUND	P provides relevant information for this domain.	This is often referred to as incorporating deterministic closure (Dörre, 1993).
MOTIVATION	P illustrates need for data, goals, methods, etc.	As shown in Meurers (1994), this is a well-motivated convention [...]
USES	Uses data, methods, etc., from P .	The head words can be automatically extracted [...] in the manner described by Magerman (1994).
EXTENSION	Extends P's data, methods, etc.	[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences to P.	Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]
FUTURE	P is a potential avenue for future work.	[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992).

Figura 1.2: Le sei etichette ACL-ARC per un articolo P

Per l'addestramento del modello è stato utilizzato un dataset di intenti di citazioni basato su un campione di articoli dell'ACL Anthology Reference Corpus [11] che include 1.941 istanze di citazioni da 186 articoli. Le citazioni sono state annotate da esperti di dominio nel campo della PNL ai quali è stato chiesto di scegliere tra le sei funzioni proposte dal classificatore ACL-ARC.

SciCite

La Allen Institute for Artificial Intelligence propone un framework di apprendimento multitasking neurale per la classificazione funzioni citazionali [12]. Più in dettaglio, i ricercatori hanno proposto un modello di deep learning che riceve in input il testo di una frase contenente una citazione. La frase viene quindi suddivisa in token per i quali le rappresentazioni vettoriali sono prodotte concatenando le rappresentazioni di parole ottenute con GloVe e le embedding ottenute con ELMo. La sequenza di vettori risultante viene inviata all'ingresso dell'LSTM bidirezionale (bi-LSTM) [13]. Quindi l'output di bi-LSTM,anch'esso una sequenza di vettori ma di dimensioni inferiori, viene inviato all'input dello strato di attenzione, il cui output è un singolo vettore. Una caratteristica importante di questo modello è l'uso di compiti aggiuntivi durante la formazione, vale a dire la sezione di classificazione e il valore della citazione. Il primo compito riguarda la sezione di previsione, in cui si è verificata la citazione, e il secondo compito riguarda l'indicazione se una frase necessita di una citazione. Pertanto, il vettore ottenuto dallo strato di attenzione viene inviato all'input di tre diversi livelli MLP, uno per ogni attività. Gli output di ciascuno di essi dopo l'attivazione di *softmax* vengono utilizzati per classificare le attività corrispondenti. Riporto in figura 1.3 lo schema di questo modello.

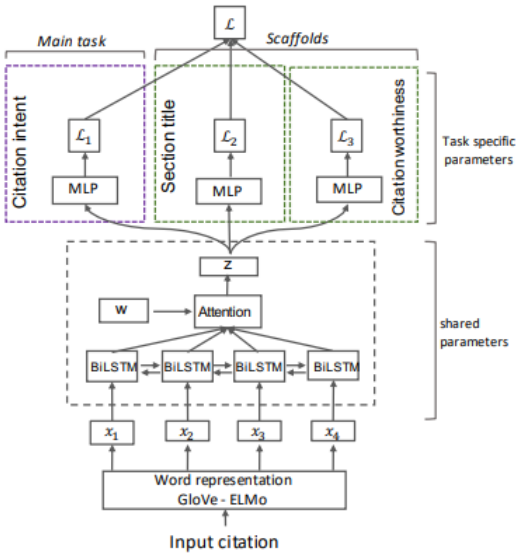


Figura 1.3: Modello di classificazione SciCite

Per l'addestramento del modello è stato introdotto un nuovo dataset di intenti di citazione che copre una varietà di domini scientifici per poter addestrare il modello e ottenere una classificazione automatica sulla citazione. SciCite fornisce uno schema di classificazione più conciso rispetto ad ACL-ARC infatti utilizza solo tre categorie di intenti (fig. 1.4): "Method", "Result Comparison" e "Background".

Intent cateogry	Definition	Example
Background information	The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field.	Recent evidence suggests that co-occurring alexithymia may explain deficits [12]. Locally high-temperature melting regions can act as permanent termination sites [6-9]. One line of work is focused on changing the objective function (Mao et al., 2016).
Method	Making use of a method, tool, approach or dataset	Fold differences were calculated by a mathematical model described in [4]. We use Orthogonal Initialization (Saxe et al., 2014)
Result comparison	Comparison of the paper's results/findings with the results/findings of other work	Weighted measurements were superior to T2-weighted contrast imaging which was in accordance with former studies [25-27] Similar results to our study were reported in the study of Lee et al (2010).

Figura 1.4: Le etichette del classificatore SciCite

Gli intenti delle citazioni sono stati annotati tramite la piattaforma di *crowdsourcing* Figure Eight. É stato selezionato un campione di documenti dal corpus di Semantic Scholar, costituito da documenti nei domini generali dell'informatica e della medicina. Agli annotatori è stato chiesto di identificare l'intento di una citazione e di selezionare tra le tre classi di funzioni. In tutto sono state annotate 11.020 citazioni provenienti da 6.627 documenti.

1.5 Aspetti da considerare in quest'area di ricerca

L'analisi delle funzioni citazionali è una metodologia utilizzata per comprendere i motivi per cui gli autori facciano determinate citazioni nelle loro pubblicazioni. Tuttavia l'area di ricerca che si occupa di studiare l'analisi di funzioni citazionali è molto eterogenea e presenta quindi alcune criticità e peculiarità da considerare.

Lo studio delle funzioni citazionali si basa su un altro ambito di ricerca, ovvero quello dell'estrazione delle citazioni, che però non è obiettivo di questa tesi.
Il numero di pubblicazioni accademiche è stimato raddoppiare ogni nove anni, è quindi necessario disporre di strumenti che consentano ai ricercatori di avere un accesso illimitato alle citazioni e poterle estrarre in modo efficiente.
Rappresentare adeguatamente le reti citazionali può essere difficile per due motivi principali: le citazione di solito non sono liberamente accessibili perché i documenti sono spesso soggetti a licenze e sono difficili da estrarre e da strutturare in modo che siano leggibili da una macchina.

In questo contesto sono nate delle iniziative tra cui I4OC[2] : una collaborazione tra editori accademici, ricercatori e altre parti interessate per promuovere la disponibilità illimitata di dati sulle citazioni accademiche. Lo scopo di questa iniziativa è promuovere la disponibilità di dati sulle citazioni in modo che siano:

- strutturati, per renderli leggibili e accessibili da una macchina in modo programmatico;
- separabili, ovvero la possibilità di analizzarli senza la necessità di accedere ai prodotti bibliografici di origine;
- aperti, ovvero accessibili senza restrizioni.

OpenCitations[3] è uno dei principali progetti dell'iniziativa IO4C. Il progetto prevede la creazione di un grande database di citazioni aperto e interconnesso, in cui ogni citazione è descritta da un insieme di metadati strutturati, come autore, titolo del documento citante, titolo del documento citato, data di pubblicazione, identificativo del DOI (Digital Object Identifier). OpenCitations ha come obiettivo la creazione di una rete di citazioni pubbliche per esplorare le connessioni tra i campi della conoscenza e per seguire l'evoluzione delle idee e delle discipline accademiche.

Un utilizzo di questa rete citazionale è quindi lo studio delle funzioni che collegano le diverse citazioni. Dopo aver dato una breve introduzione di questo contesto possiamo approfondire quelli che sono i principali aspetti da considerare nell'ambito di ricerca dell'analisi delle funzioni citazionali.

Eterogeneità dei campioni
L'analisi delle funzioni citazionali può essere condotta su una vasta gamma di pubblicazioni, queste possono appartenere a ambiti di ricerca differenti. Un aspetto da considerare nell'analisi è quindi l'eterogeneità dei possibili ambiti da cui provengono

[2]IO4C: https://i4oc.org/
[3]OpenCitations: https://opencitations.net/about

le citazioni. Ogni ambito ha le proprie caratteristiche e potrebbe essere interessante sapere quali differenze portano allo studio delle funzioni citazionali.
In questa tesi teniamo conto anche di questo aspetto tenendo traccia della disciplina di origine delle citazioni.

Scarsità dei dati annotati
Una problematica che può riscontrarsi nella analisi delle funzioni citazionali è la scarsità dei dati annotati manualmente. Il processo di annotazione manuale è lungo e richiede diverse opinioni da parte di ricercatori qualificati nell'ambito delle pubblicazioni considerate. La scarsità di dati annotati può limitare la capacità di addestrare e validare i modelli e portare a una performance insufficiente nella classificazione delle funzioni citazionali. Se si vuole utilizzare un classificatore che utilizza tecniche di machine learning, più dati dati si raccolgono più il meccanismo di apprendimento supervisionato funziona.

Divergenze linguistiche
L'analisi può essere influenzata anche dalla lingua in cui è scritto l'articolo originale. Ad esempio, alcune lingue potrebbero avere una maggiore propensione all'utilizzo di determinate funzioni citazionali rispetto ad altre. Ciò può rendere difficile confrontare i risultati tra studi condotti su articoli scritti in lingue diverse.

Differenza tra l'etichettatura di funzioni citazionali
Un altro apsetto da considerare è la differenza nell'etichettatura delle funzioni citazionali tra i diversi classificatori. Questa eterogeneità deriva dai campi di applicazioni dei classificatori, infatti possono essere progettati per diversi casi di studio e in base alle singole esigenze di sviluppo si preferisce usare delle specifiche etichette. Ad esempio, come visto sopra, le classi del classificatore ACL-ARC sono sei ("background", "motivation, "uses", "extension", "comparison or contrast", "future"), mentre quelle di SciCite sono solo tre ("background information", "method", "result comparison").
Per questo motivo non è possibile fare un confronto diretto ed è difficile confrontare i risultati tra diversi studi. É però possibile cercare i collegamenti tra le etichette e vedere se ci sono delle parziali o totali sovrapposizioni nella funzione citazionale che eprimono, questo sarà uno degli argomenti trattati in questa tesi.

Difficoltà nell'annotazione manuale
Non solo i classificatori ottengono risultati diversi, ma anche gli annotatori umani come dimostrato nell'esperimento condotto da A. Di Iorio [14]. L'obiettivo dell'esperimento è indagare come un modello di riferimento esistente per classificare le citazioni, in questo caso CiTO (Citation Typing Ontology), è interpretato e utilizzato dagli annotatori della letteratura scientifica.
CiTO definisce quarantuno proprietà che consentono agli utenti di caratterizzare con precisione lo scopo di una citazione. Al momento non sembra esistere un'altra ontologia OWL (Web Ontology Language) che fornisca un insieme di proprietà per annotare i tipi di citazione così ricco. Questo aspetto ha contribuito all'adozione dell'ontologia da parte della comunità di Semantic Publishing, ma d'altra parte, questa ricchezza è percepita come un ostacolo da alcuni annotatori.

L'esperimento ha coinvolto due gruppi, ciascuno composto da dieci soggetti ai quali è stato chiesto di assegnare una delle proprietà di CiTo a 105 citazioni. Il primo gruppo ha utilizzato l'elenco completo di 41 proprietà CiTO. Invece, il secondo ha svolto lo stesso compito utilizzando solo 10 proprietà CiTO. I risultati hanno evidenziato una grande varietà nelle scelte degli esseri umani. Infatti, solo 18 citazioni nel primo gruppo e 24 citazioni nel secondo sono state classificate con esattamente la stessa proprietà CiTO da almeno 6 soggetti. È stato utilizzato il Fleiss' k per valutare l' affidabilità dell'accordo tra le scelte degli annotatori, evidenziando che entrambi i gruppi avevano un accordo basso ma nel secondo l'indice era più alto rispetto al primo. Al termine dell'esperimento, a entrambi i gruppi è stato chiesto di rispondere a un questionario SUS (System Usability Scale) per dare un punteggio di usabilità a CiTo: è emerso un valore più alto dal gruppo con meno funzioni citazionali.
Questo studio mette in evidenza che anche il processo di annotazione manuale può risultare difficile e può portare a risultati diversi, in particolare quando un classificatore utilizza tante proprietà per esprimere l'intento di una citazione.

Capitolo 2

Progettazione dell'applicazione

2.1 Obiettivo del progetto

Come descritto nel capitolo precedente, esistono al momento tanti studi che portano a risultati differenti, principalmente per i seguenti motivi:

- alcuni modelli di classificazione di citazioni sono addestrati su pochi dati target
- i dati target riguardano spesso ambiti diversi
- il processo di annotazione manuale risulta lungo e complicato
- i classificatori utilizzano diverse etichette

Questa tesi nasce dalla necessità di creare un ambiente di annotazione manuale che possa produrre dei dati validi per aiutare a migliorare i classificatori esistenti o per addestrarne di nuovi.
La domanda che bisogna porsi è quindi: "Come possiamo risolvere i problemi di annotazione e proseguire nella ricerca ottenendo dei dati accurati e affidabili in modo rapido e semplice?"

Attualmente esistono già delle piattaforme di *crowdsourcing* dove gli utenti possono annotare manualmente. A questi viene fornita una lista di citazioni e gli viene chiesto di esprimere una loro opinione sulla funzione che svolgono. Questa scelta è vincolata alle etichette del classificatore che è oggetto di studio. L'obiettivo di queste piattaforme è raccogliere dati necessari per l'addestramento del classificatore.
Si pensa ora ad un modo per migliorare questo processo di annotazione manuale permettendo a chi deve svolgere questo lavoro anche la possibilità di confrontare le etichette estratte da più classificatori. Ricordiamo che due classificatori possono ottenere sulla stessa citazione una funzione differente oppure identificare lo stesso scopo ma rappresentarlo con etichette diverse.
Ad esempio, prendiamo la seguente citazione:

"For the learning-to-rank method, we used the source code of [39]".

Considerando ACL-ARC(1.2) una etichetta adatta a questo scopo potrebbe essere "uses" mentre per SciCite(1.4) potrebbe essere "method".

Da queste sostanziali differenze tra le etichette dei classificatori emerge la necessità

di evidenziare il livello di *agreement/disagreement* tra di essi per poter migliorare l'accuratezza della classificazione manuale delle funzioni citazionali. Si vuole quindi evidenziare l'accordo che ci può essere tra le predizioni di più classificatori e mostrarle all'annotatore nel momento in cui deve fare la propria scelta. In questo modo guardando la stessa citazione può vedere come si comportano altri classificatori e in base al loro livello di accordo esprimere una propria opinione.

L'obiettivo finale è quindi di creare un ambiente per l'annotazione manuale, permettendo di confrontare diversi classificatori e di costruire un ambiente per produrre dati accurati.
Questo ambiente può essere sviluppato mediante una applicazione web e deve quindi fornire specifiche funzionalità.

- Deve permettere agli utenti di annotare manualmente su una collezione di citazioni fornite dal sistema.
- Deve essere possibile utilizzare le etichette di uno o più classificatori per assegnare una funzione ad una citazione.
- Deve mostrare, su una citazione, tutte le etichette predette dai classificatori evidenziando il loro *agreement*.
- Anche nel caso in cui non fossero presenti delle predizioni l'utente può sempre assegnare l'etichetta che ritiene più adatta.
- Deve permettere di distinguere le citazioni in base alla disciplina o il periodo di pubblicazione o altri criteri richiesti dall'utente.

L'applicazione che si ha intenzione di realizzare può avere più casi d'uso. Un primo caso d'uso è annotare a partire dalle etichette e dalle predizioni di un solo classificatore. Un altro utilizzo intuitivo è selezionare funzioni citazionali a partire dalle diverse predizioni dei differenti classificatori e vedere i livelli di *agreement* tra di loro. Può anche essere sfruttato per studiare il progresso di uno specifico classificatore. Mettendo in evidenza il livello di *agreement* nelle predizioni tra due versioni dello stesso classificatore può essere visualizzato il miglioramento o il peggioramento del suddetto. L'utente trovandosi d'accordo con la predizione della nuova versione del classificatore, per esempio, può affermare che ci sia stato un miglioramento. Un altro caso d'uso, che non sfrutta però totalmente il potenziale dell'applicazione, è l'annotazione senza predizioni di alcun classificatore, che come già detto deve sempre essere garantito.

2.2 I requisiti

Vediamo in questa sezione una serie di requisiti che deve avere l'applicazione per raggiungere l'obiettivo finale del progetto.

Un ambiente per la produzione di dati annotati
Si vuole creare un ambiente distribuito che permetta ai ricercatori di annotare manualmente in modo semplice e rapido. Si pensa quindi ad un ambiente dove tutti i

ricercatori possono accedere e utilizzare i dati presenti nella piattaforma per annotare manualmente e velocizzare questo processo. A questo scopo è necessario che il sistema riesca a memorizzare migliaia di annotazioni manuali.
L'applicazione deve essere anche accessibile. Per ricreare questo ambiente si pensa quindi ad un'applicazione Web distribuita per mezzo di un network, come ad esempio una Intranet all'interno di un sistema informatico o attraverso internet, cioè in un'architettura tipica di tipo *client-server*.
Agli utenti devono essere fornite delle credenziali per accedere ed iniziare ad annotare delle citazioni già fornite dal sistema.

Il "Mapping" tra le etichette dei classificatori
L'applicazione deve permettere di integrare dei classificatori insieme alle loro etichette, gli utenti devono poterle visualizzare e selezionare per annotare una citazione.
La semplice installazione di più classificatori non è sufficiente al nostro obiettivo anzi questo potrebbe portare solo confusione al processo di annotazione manuale, proprio perché, come già detto, non esiste una standardizzazione esplicita che unisce le diverse etichette.
Ogni classificatore ha un proprio set di etichette per fare predizioni su una citazione.Ad esempio, ACL-ARC utilizza sei etichette, mentre SciCite ne utilizza solo tre. Questo però non significa che le funzioni che esprimono siano diverse ma che semplicemente sono rappresentate con nomi diversi.

Da qui nasce un requisito fondamentale dell'applicazione: la possibilità di avere un *mapping* esplicito tra le possibili etichette dei diversi classificatori. Questo avviene instaurando una corrispondenza tra le etichette di diversi classificatori che non è necessariamente 1:1, ma in base all'affinità delle funzioni che esprimono.

Riporto di seguito in figura 2.1 un esempio di *mapping* tra i classificatori ACL-ARC, SciCite e SCAR.

<table>
<tr><th>ACL-ARC</th><th>SciCite</th><th>SCAR</th></tr>
<tr><td>Extends</td><td rowspan="4">Background</td><td>Extends</td></tr>
<tr><td>Background</td><td rowspan="3">Cites</td></tr>
<tr><td>Motivation</td></tr>
<tr><td>Future</td></tr>
<tr><td>Compare</td><td>Compare result</td></tr>
<tr><td rowspan="2">Uses</td><td rowspan="2">Method</td><td>Uses method</td></tr>
<tr><td>Uses data from</td></tr>
</table>

Figura 2.1: Esempio di mapping tra ACL-ARC,SciCite e SCAR

Come possiamo vedere in questo esempio di *mapping*, l'etichetta "Background" di SciCite corrisponde a quattro diverse etichette di ACL-ARC ("Extends", "Background", "Motivation", "Future") e a due di SCAR ("Extends", "Cites").

Il *mapping* deve essere impostato a priori manualmente e deve esserci la possibilità di cambiarlo.

Il meccanismo di "Agreement"
Per velocizzare l'annotazione manuale si pensa alla possibilità di avere già delle predizioni per ogni citazione e mostrare il loro livello di *agreement*. L'idea è la seguente: presa una qualsiasi citazione è possibile vedere graficamente quali etichette estraggono automaticamente i classificatori e quale livello di accordo c'è tra di esse.
In questo modo il lavoro di un annotatore è semplificato ulteriormente, infatti può decidere se approvare direttamente ciò che identificano le predizioni automatiche oppure scegliere un'etichetta diversa.

Facciamo degli esempi su come un utente dovrebbe visualizzare *agreement* data una citazione e le predizioni dei tre classificatori ACL-ARC, SciCite e SCAR.

1) Caso Agreement totale: tutte le tre predizioni rappresentano la stessa funzione citazionale.

Citazione	The MSRA10K dataset [53] is composed of 10,000 images randomly selected from the MSRA dataset.

Classificatore	**Predizione**
ACL-ARC	uses
SciCite	method
SCAR	uses data from

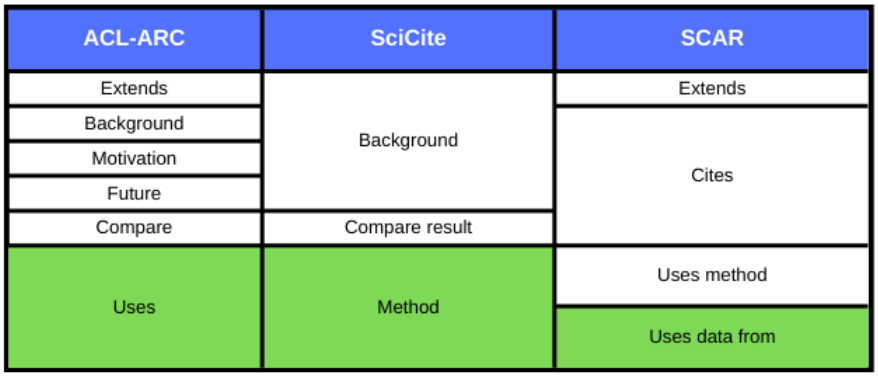


Figura 2.2: Esempio agreement totale tra ACL-ARC, SciCite e SCAR

2) Caso Agreement parziale: esiste almeno una predizione che ha una funzione citazionale diversa dalle altre.

Citazione	Finally, the BPSK waveforms which realise this covariance matrix are generated as follows [13] where X is the transmit waveforms matrix and is a matrix including zero mean and unit variance Gaussian random variables.

Classificatore	**Predizione**
ACL-ARC	uses
SciCite	background
SCAR	uses method

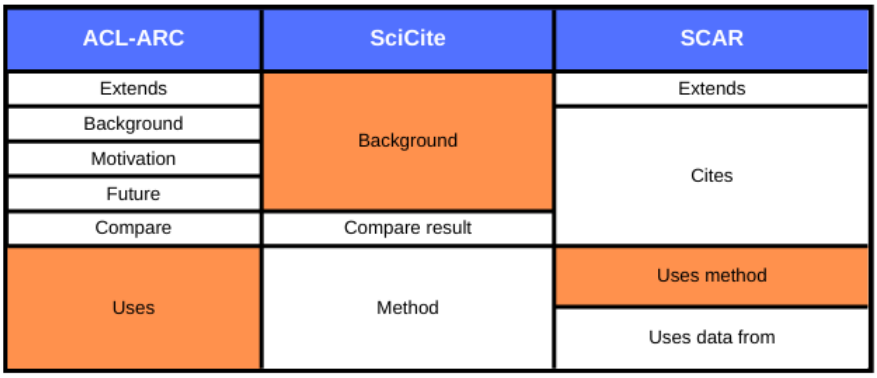


Figura 2.3: Esempio agreement parziale tra ACL-ARC, SciCite e SCAR

3) Caso Agreement nullo: tutte le tre predizioni hanno una funzione citazionale diversa.

Citazione	We participated in the 2015 Ischemic Stroke Lesion Segmentation (ISLES) challenge, where our system achieved the best results among all participants on subacute ischemic stroke lesions (Maier et al., 2017).

Classificatore	**Predizione**
ACL-ARC	compare
SciCite	background
SCAR	uses method

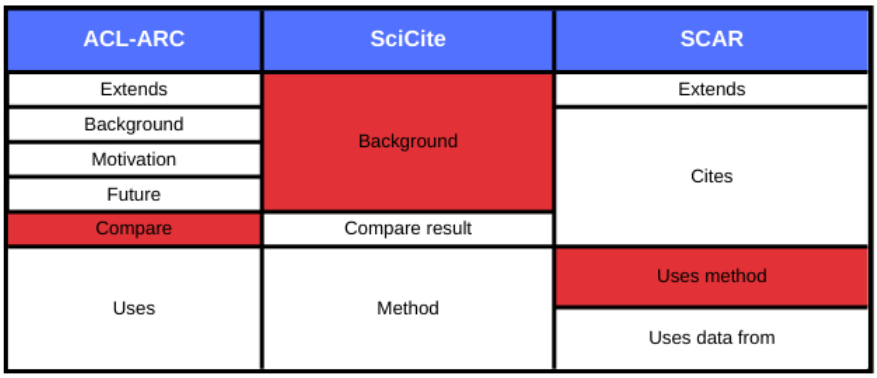


Figura 2.4: Esempio agreement nullo tra ACL-ARC, SciCite e SCAR

La visualizzazione così intuitiva dell'*agreement* tra le etichette dei classificatori rende semplice per l'utente l'utilizzo dell'applicazione per i possibili casi d'uso.

Gestione dei classificatori
L'applicazione, come già specificato, deve ospitare al suo interno più classificatori. Questi però non possono essere fissi, perché renderebbe l'applicazione poco flessibile. Se è necessario produrre dei dati annotati per un nuovo classificatore deve esserci la possibilità di introdurlo, viceversa se non si ritiene più valido un classificatore, quest'ultimo deve essere rimosso. Visto che la ricerca delle funzioni citazionali è in continua evoluzione una gestione dei classificatori è utile anche per stare al passo con i tempi, per esempio integrando i classificatori emergenti. Deve esserci quindi la possibilità in qualsiasi momento di togliere un classificatore oppure aggiungerne uno e fare in modo che l'applicazione si adatti per annotare su quest'ultimo.

Per produrre dati più accurati, il modo migliore è annotare utilizzando tutti i classificatori così che tra un più vasto range di opzioni venga scelta l'etichetta del che rappresenta la funzione citazionale più adatta per una determinata citazione.
Inoltre se abbiamo più classificatori, tutte le loro predizioni e i dati annotati, è possibile identificare il classificatore migliore.

Fonte espandibile di citazioni
Per aumentare la quantità di citazioni che fornisce l'applicazione deve esserci un meccanismo che permette di importarne di nuove. Questa funzionalità però non deve essere accessibile a tutti per una questione di qualità delle fonti. Si pensa quindi ad una distinzione tra utenti e amministratori, di cui solo gli ultimi possono aggiungere nuove citazioni, per esempio facendo un *upload* di un semplice file.

Categorizzazione delle citazioni
Le citazioni possono provenire da documenti scientifici di ambiti diversi, per questo motivo deve esserci la possibilità di distinguerle associando ad esse una categoria. Si pensa quindi ad un modo efficace per stabilire una relazione citazione-categoria per esempio introducendola nel momento di importazione dei file.

Criteri per arricchire i dati annotati
Per arricchire la quantità di annotazioni è utile aggiungere dei criteri per la ricerca delle citazioni. L'applicazione deve caricare e fornire all'accesso delle citazioni e se l'utente ne desidera alcune specifiche deve esserci un filtro di ricerca apposito. Per esempio un utente deve essere in grado di estrarre solo le citazioni della categoria in cui è più competente.

Una lista di possibili filtri per arricchire con criterio i dati annotati è la seguente.

- categoria/ambito
- anno di pubblicazione
- limitazione sul numero di citazioni da annotare
- livello di *agreement*
- nuove citazioni non ancora annotate
- selezione dei classificatori da considerare

In alcuni momenti potrebbe essere necessario ottenere dei dati annotati solo su alcune specifiche citazioni, a questo scopo l'applicazione può forzare tutti gli utenti a ad annotare la stessa lista di citazioni. Questa funzionalità, come l'importazione di nuovi dati, deve essere accessibile solo agli amministratori.

API per il flusso di dati
Un altro tassello importante riguarda la comunicazione *client-server*. Dagli obiettivi dell'applicazione possiamo dedurre due flussi principali di dati, uno per le annotazioni e uno per le citazioni. Un utente deve poter ottenere le citazioni che vuole annotare secondo i criteri richiesti e deve poter inviare le proprie annotazioni.
Per garantire ciò, l'applicazione deve esporre una propria API (Application Program Interface), consentendo al client e server di comunicare tra loro tramite richieste e risposte HTTP. L'applicazione deve fornire un insieme di endpoint (punti di accesso), dove ognuno di questi deve rispondere ad una specifica richiesta e dare una risposta adeguata.

Capitolo 3

Panoramica del Citation Intent Trove (CIT)

Partendo dai requisiti descritti nel capitolo precedente abbiamo creato il Citation Intent Trove (CIT) un ambiente web per l'annotazione manuale di citazioni.
In questo capitolo vedremo come può essere utilizzata l'applicazione mostrando l'interfaccia grafica e le possibili interazioni con essa.
L'utilizzo del CIT si distingue in base al tipo di utente, che può essere un annotatore o un amministratore. Dividiamo in due sezioni le varie funzionalità offerte dall'applicazione in base al tipo di utente.
Ricordiamo che è possibile accedere solo tramite l'utilizzo di credenziali valide fornite da un amministratore.

3.1 Utente Annotatore

Il primo passo è utilizzare un browser e recarsi nel dominio in cui è stato effettuato il deploy del CIT, in seguito viene caricata la pagina iniziale di login dove bisogna inserire le proprie credenziali.

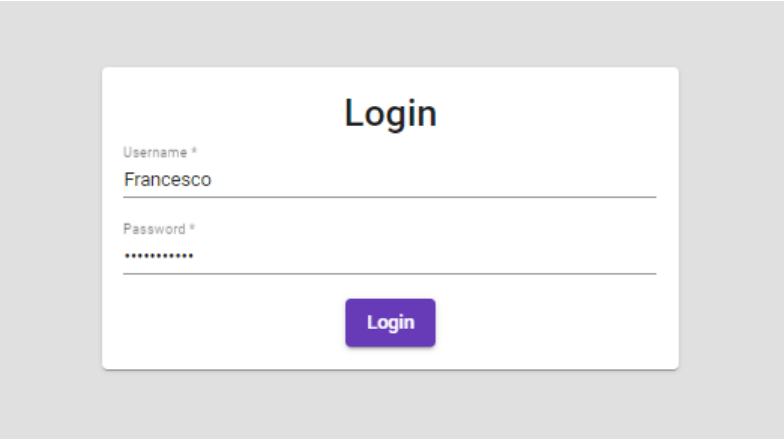


Figura 3.1: Login page

Una volta effettuato il login, entriamo nel vero e proprio workspace dove si possono visualizzare e annotare le citazioni.

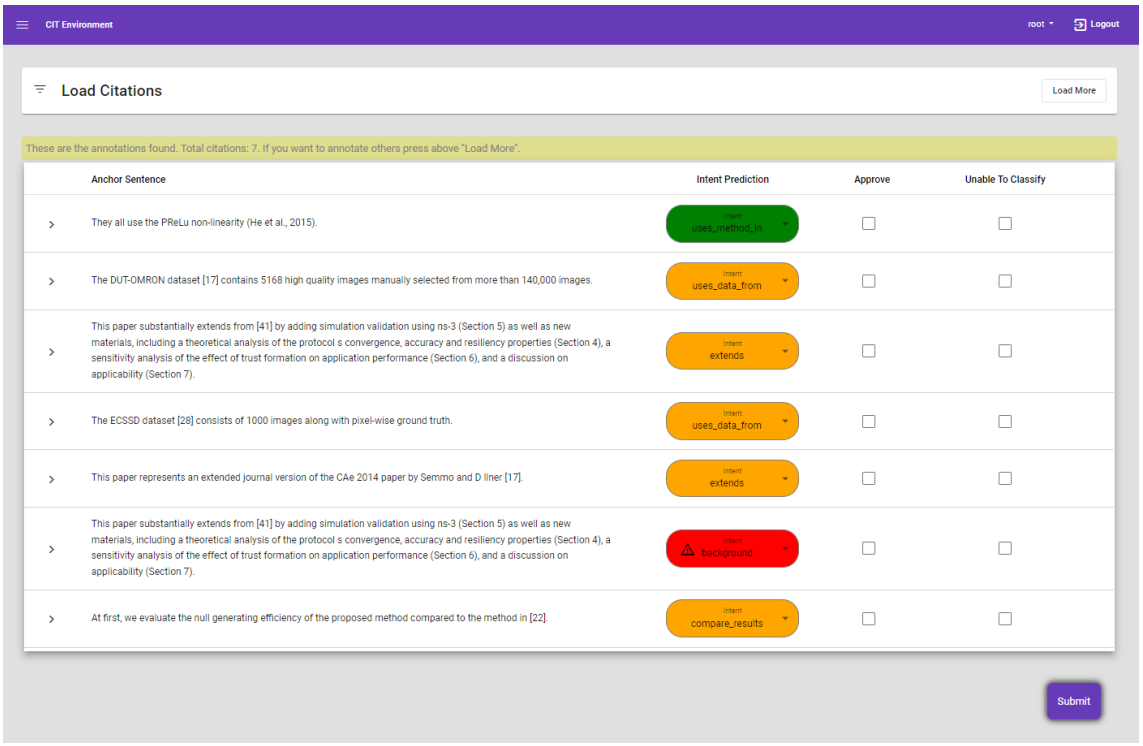


Figura 3.2: Overview UI del Citation Intent Trove

Il workspace è composto da:

- un header, tramite il quale si può effettuare il logout oppure cambiare la password.
- i filtri di ricerca delle citazioni
- la lista di citazioni trovate in base ai filtri

Al primo accesso viene caricato il primo modulo contenente 20 citazioni randomiche che ancora non sono state annotate. Se si vuole annotare delle specifiche citazioni basta cliccare su "Load More" e utilizzare i filtri proposti.

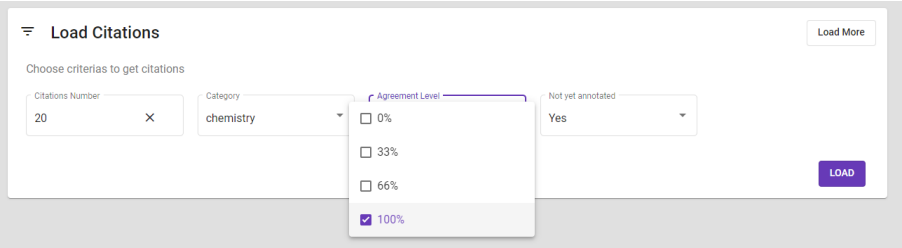
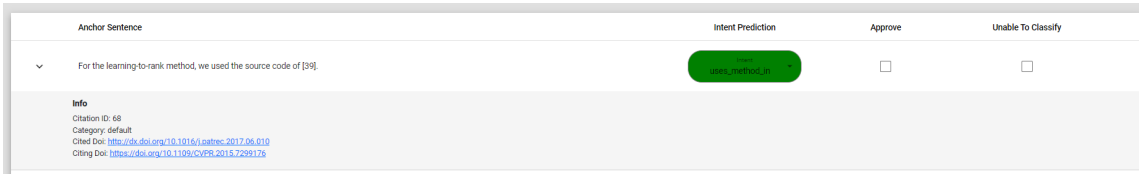
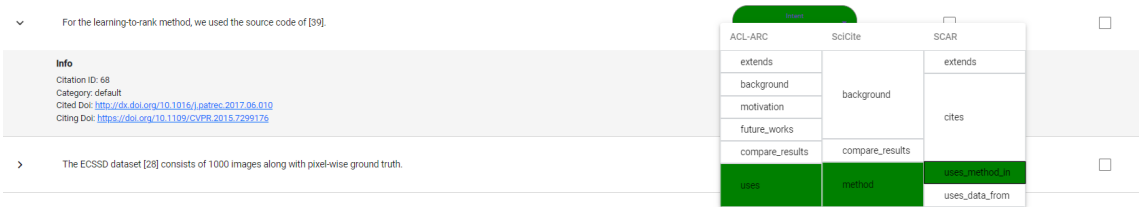


Figura 3.3: Filtri di ricerca delle citazioni

Guardiamo ora da più vicino il processo di annotazione. Consideriamo una singola riga della lista:



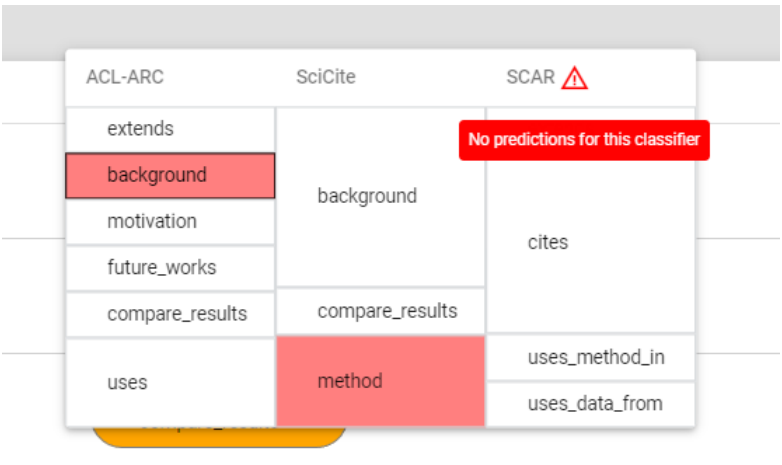
Come possiamo vedere dall'immagine viene mostrato il testo e tutte le informazioni associate per quella citazione.
Nella colonna "Intent Prediction" viene suggerita la predizione considerata più accurata. Cliccandoci sopra si possono visualizzare tutte le predizioni per ogni classificatore con il loro mapping e livello di agreement.



Il colore delle etichette predette dai classificatori cambiano in base al livello di agreement.

Livello agreement	Colore prediction
0%	rosso
$> 0\%$ & $\leq 50\%$	arancione
$> 50\%$ & $\leq 99\%$	giallo
100%	verde

Esistono dei casi in cui una citazione non ha delle predizioni per alcuni classificatori, in questo caso viene mostrato un warning.



L'utente a questo punto, presa una citazione dalla lista, può fare 4 scelte:

1. approvare l'etichetta suggerita cliccando sulla casella "approve";
2. cambiare l'etichetta cliccando su quella che ritiene più adatta alla citazione;
3. segnalare un problema clicando sulla casella "Unable To Classify" e selezionando una motivazione;
4. non eseguire nessuna operazione.

Dopo aver fatto questa scelta per tutte le citazioni, può inviare le sue annotazioni tramite il pulsante "submit".

3.2 Utente Amministratore

L'utente amministratore è il tipo di utente che ha accesso a tutte le funzionalità dell'applicazione, quindi a tutte quelle di un annotatore più alcune specifiche.
Le funzionalità aggiuntive sono:

- la gestione degli utenti;
- la gestione del filtro di default;
- l'importazione di nuove citazioni

Gestione degli utenti Tramite questa interfaccia è possibile:

- visualizzare tutti gli utenti che hanno l'autorizzazione ad accedere all'applicazione
- E' possibile resettare la password di un utente nel caso l'avesse dimenticata
- Cancellare un utente negandogli l'accesso al CIT
- Aggiungere un nuovo utente

Gestione del filtro di default Il filtro di default definisce un filtro di ricerca sulle citazioni univoco per tutti gli utenti. Quando questo è attivo gli utenti non possono fare ricerche sulle citazioni ma saranno obbligati ad annotare quelle automaticamente generate. Il filtro può essere attivato e disattivato da un amminstratore in qualsiasi momento.

Importazione di nuove citazioni Se si vuole arricchire il database di nuove citazioni è possibile farlo semplicemente caricando un file in formato csv. Per la corretta importazione bisogna adattare le colonne del file con quelle che l'applicazione si aspetta (visibili ad interfaccia), ricordandosi che:

- Le colonne possono essere in qualsiasi ordine;
- Se non esiste una colonna di quelle elencate, il campo della citazione sarà vuoto, ad eccezione del `citation_id`;
- Se il `citation_id` non esiste, verrà aggiunto automaticamente;
- Se una cella è vuota verrà impostata su `NULL`;
- Le colonne non previste non vengono considerate.

Dopo aver caricato il file è necessario specificare la categoria delle citazioni che si vogliono importare scegliendo il tipo di riempimento:

- Fill all: tutte le citazioni verranno riempite con la categoria specificata sovrascrivendo un'eventuale categoria già esistente nel file;
- Only not specified: solo le citazioni che non hanno una categoria verranno riempite con la categoria specificata.

Scelta la categoria è possibile terminare la procedura di importazione cliccando sul bottone "Upload".

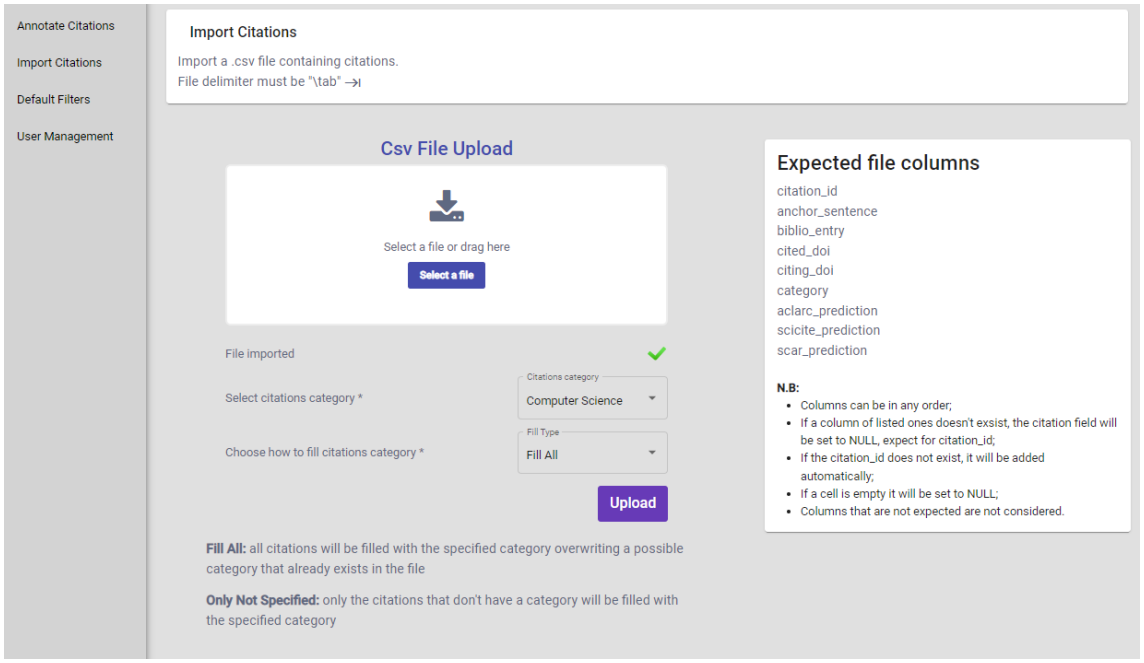


Figura 3.4: Interfaccia per l'importazione di nuove citazioni

Capitolo 4

Implementazione Citation Intent Trove (CIT)

4.1 Tecnologie utilizzate

Per la realizzazione della applicazione web sono state utilizzate alcune delle tecnologie più recenti del momento. Per lo sviluppo back-end parliamo di Node.js insieme al framework Express mentre per lo sviluppo front-end del framework Angular. L'IDE utilizzato è Visual Studio Code.

Node.js è un ambiente di runtime multipiattaforma open source che consente agli sviluppatori di creare tutti i tipi di strumenti e applicazioni lato server in JavaScript. Il runtime è destinato all'uso al di fuori di un contesto del browser (ovvero eseguito direttamente su un computer o sistema operativo del server). Pertanto, l'ambiente omette le API JavaScript specifiche del browser e aggiunge il supporto per le API del sistema operativo più tradizionali, incluse le librerie HTTP e file system. Node, dal punto di vista dello sviluppo di un server Web, presenta una serie di vantaggi:

- E' progettato per ottimizzare il throughput e la scalabilità nelle applicazioni Web;
- Il codice è scritto in JavaScript, il che significa che si impiega meno tempo a gestire il "context switch" tra i linguaggi quando si scrive sia codice lato client che lato server;
- Permette di convertire altri linguaggi in JavaScript, ed è quindi possibile utilizzare linguaggi come TypeScript, CoffeeScript, ClojureScript, Scala, LiveScript, ecc;
- Fornisce l'accesso a centinaia di migliaia di pacchetti riutilizzabili tramite il package manager "npm";
- E' portatile infatti è disponibile su Microsoft Windows, macOS, Linux, Solaris, FreeBSD, OpenBSD, WebOS e NonStop OS. Inoltre, è ben supportato da molti provider di web hosting, che spesso forniscono infrastrutture e documentazione specifiche per l'hosting di siti Node.
- Ha un ecosistema di terze parti molto attivo e una comunità di sviluppatori, con molte persone disposte ad aiutare.

Express è un framework open-source Node.js per la programmazione di applicazioni web e mobile. Minimalista, leggero e veloce, Express potenzia Node.js senza comprometterne le funzionalità.
Il framework Express consente di creare potenti API di routing e utilizzare dei middleware per rispondere alle richieste HTTP tramite semplici meccanismi.

Angular è un framework front-end JavaScript open source scritto in TypeScript. Google lo mantiene e il suo scopo principale è sviluppare applicazioni single page. Come framework, Angular fornisce una struttura standard con cui gli sviluppatori possono lavorare e consente di creare applicazioni di grandi dimensioni in modo gestibile. Un'applicazione Angular può in genere essere eseguita su tutti i browser (ad esempio: Chrome, Firefox) e sistemi operativi, come Windows, macOS e Linux.

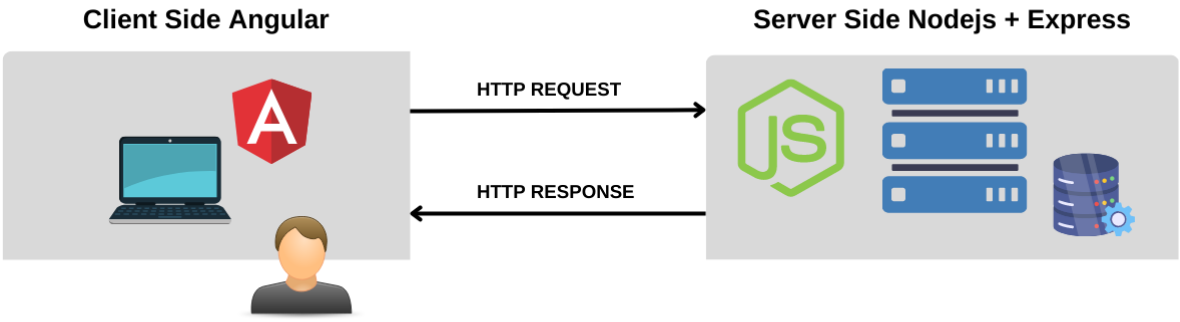


Figura 4.1: Struttura Client e Server side della applicazione

4.2 Struttura Backend

Il lato server racchiude tutto ciò che è necessario per far funzionare l'applicazione, ma che gli utenti non vedono e non interagiscono direttamente. In questo caso una parte server-side è necessaria per la costruzione ed esposizione di un API REST con la quale il front-end (Angular) può comunicare, predisponendo la base per la lettura dei dati ricevuti in input e il loro salvataggio nel database. L'applicazione web rimane in ascolto di richieste HTTP e per ogni client crea una sessione univoca. L'interazione avviene definendo un meccanismo di routing ovvero al modo in cui gli endpoint (URI) dell'app devono rispondere alle richieste del client.

In questo caso l'applicazione definisce un percorso e un flusso di codice standard per gestire ogni richiesta, in modo da avere un codice di qualità. Infatti il routing è composto da un insieme di controller ognuno suddiviso per entità e con ognuno con un suo manager. Il controller specifica ogni route disponibile per una determinata entità e delega la funzione associata al suo manager.

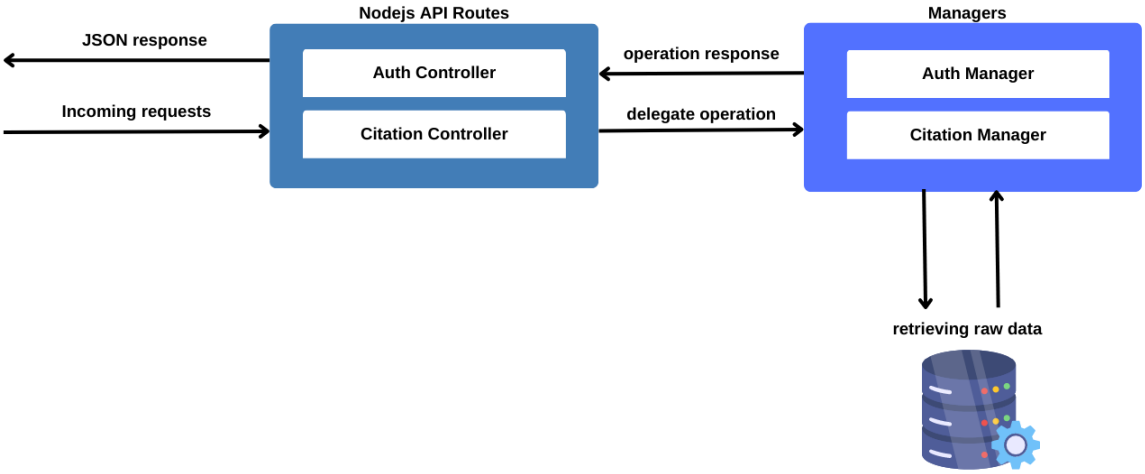


Figura 4.2: Routing server NodeJs

Una gestione di questo tipo garantisce una buona espandibilità del codice e una facile manutenibilità.
Ho deciso di manipolare i dati tramite formato JSON, perfetto per questo tipo di sviluppo.

Quali chiamate REST gestisce l'applicazione? Alcuni esempi di gestione sono:

- la ricerca filtrata di citazioni

- il salvataggio di nuove citazioni
- l'update delle citazioni
- il login e logout degli utenti

4.3 Struttura database

Un componente fondamentale dell'applicazione è il database. Ho scelto MySQL un sistema di gestione di database relazionali (RDBMS) sviluppato da Oracle basato sul linguaggio di query strutturato (SQL). Questo tipo di gestione è perfetto, infatti nel dominio dell'applicazione esistono diverse entità relazionate tra loro, in questo modo posso rappresentarle facilmente.
Nell'immagine 4.3 osserviamo la struttura del database composto dalle entità:

- Citation: rappresenta una citazione.
- Annotated Citations: rappresenta una annotazione, ovvero la scelta di un utente sulla funzione che secondo lui ha una determinata citazione;
- User: rappresenta un utente autorizzato ad accedere all'applicazione;
- Citation Error: rappresenta un problema associato ad una determinata citazione;
- Default Filters: rappresenta un filtro di ricerca per le citazioni.

Guardando attentamente l'entità `Citation`, possiamo notare tra i suoi attributi le `prediction` (es. `scar_ prediction`), è in questi campi che vengono salvate le predizioni dei classificatori.

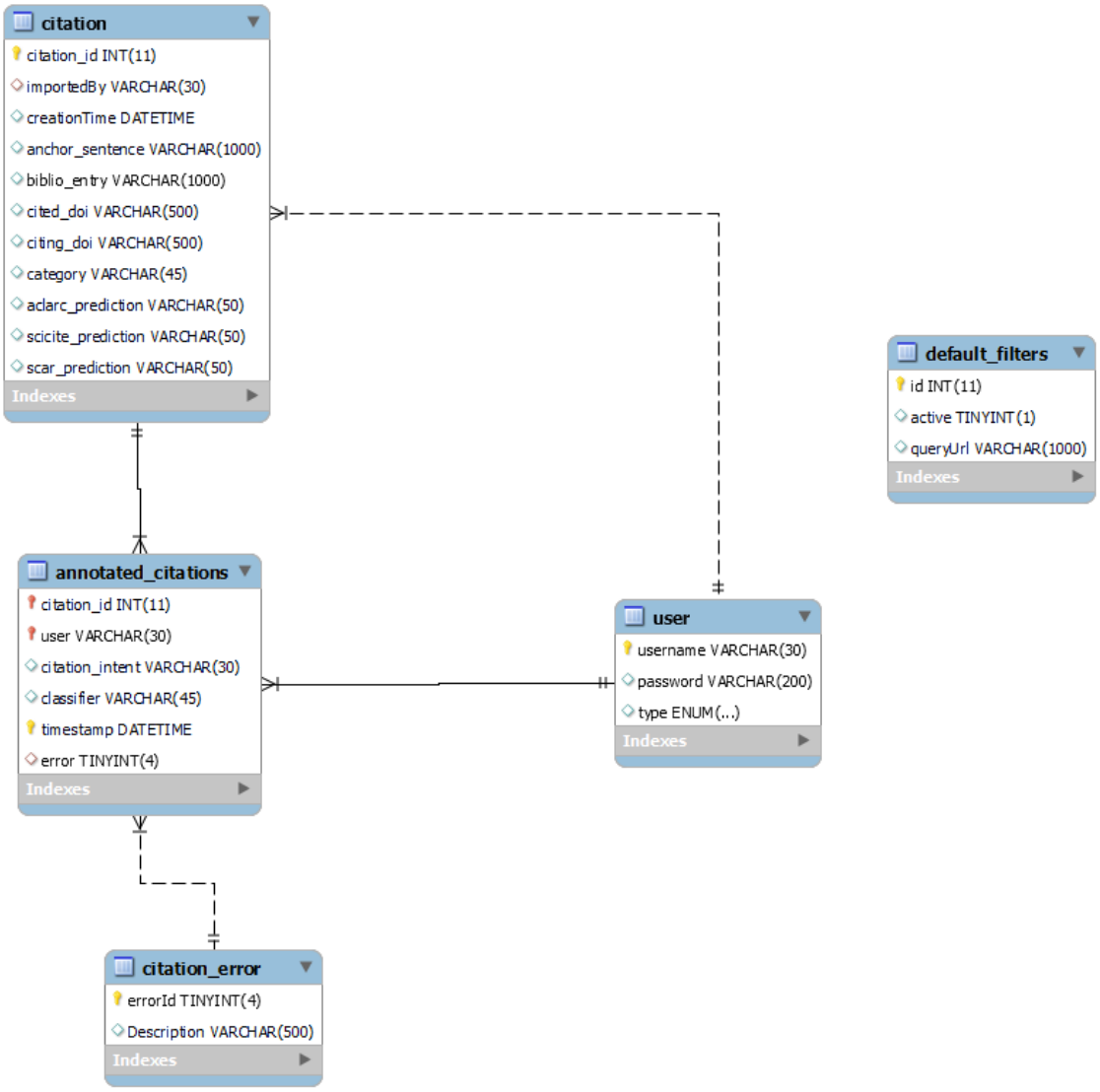


Figura 4.3: Rappresentazione struttura del database MySql

Il server Nodejs si appoggia direttamente al server MySql istanziando delle connessioni per richiedere l'esecuzione di query.

4.4 Dinamicità dell'applicazione

Il Citation Intent Trove, come già anticipato, raccoglie al suo interno più classificatori. Ma in che modo possono essere aggiunti e come fa l'applicazione a riconoscerli? Su quanti classificatori si può annotare e quante predizioni possono esserci per una singola citazione?

La progettazione dell'intera applicazione è iniziata prendendo in considerazione che in futuro esisteranno nuovi modelli di classificatori che saranno utili da integrare. Una citazione deve poter contenere infinite predizioni per infiniti classificatori e deve permettere di annotare su tutti questi. Per questo motivo è stata creata una SSOT(Single Source Of Truth) tramite la quale l'applicazione può capire in un determinato momento quali sono i classificatori da considerare.
La SSOT consiste in due file JSON uguali, uno server-side e uno client-side. Vediamo di seguito un esempio del file:

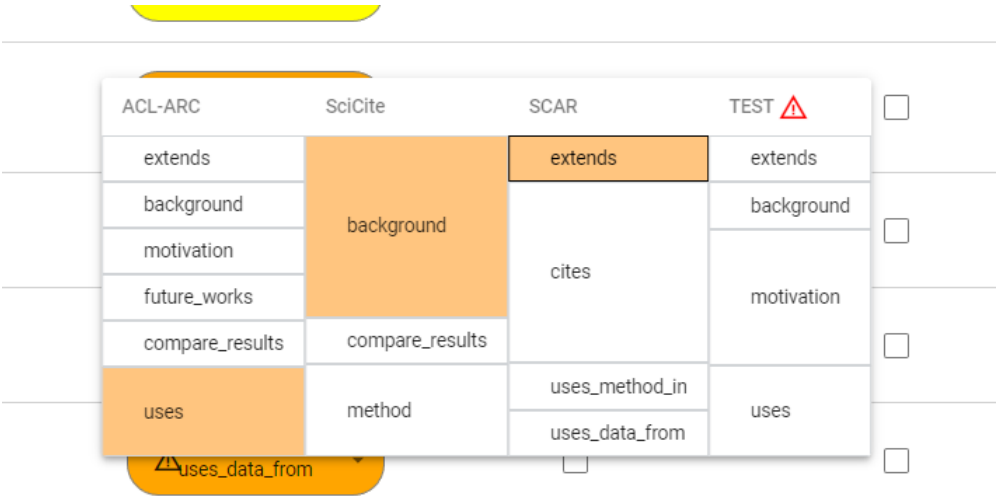
```
[{
    "classifier" : "ACL-ARC",
    "prediction": "aclarc_prediction",
    "intents" : {
        "extends" : 1,
        "background" : 1,
        "motivation" : 1,
        "future_works" : 1,
        "compare_results" : 1,
        "uses" : 2
    }
},
{

    "classifier" : "SciCite",
    "prediction": "scicite_prediction",
    "intents" : {
        "background" : 4,
        "compare_results" : 1,
        "method" : 2
    }
},
{
    "classifier" : "SCAR",
    "prediction": "scar_prediction",
    "intents" : {
        "extends" : 1,
        "cites" : 4,
        "uses_method_in" : 1,
        "uses_data_from" : 1
    }
}
]
```

Ogni singolo oggetto del file rappresenta un classificatore con il suo nome, il nome della predizione sul database e le funzioni citazionali che può assegnare. Il front-end e il back-end del Citation Intent Trove si adattano a ciò che è qui scritto, infatti una modifica in questo file comporta il riadattamento della struttura del database e la modifica dell'interfaccia utente.

Proviamo ad aggiungere un oggetto classificatore denominato "TEST":

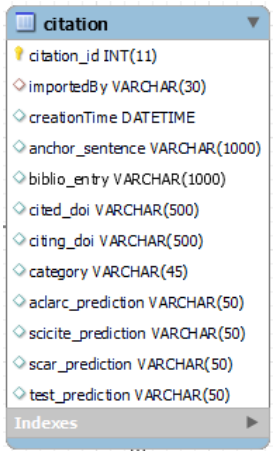
```
{
    "classifier" : "TEST",
    "prediction": "test_prediction",
    "intents" : {
        "extends" : 1,
        "backgroun" : 1,
        "motivation" : 3,
        "uses" : 2
    }
```

Osserviamo il risultato della modifica a interfaccia utente:



Come vediamo l'oggetto è stato aggiunto correttamente a interfaccia mostrando il nome del classificatore e tutte le sue etichette. Notiamo anche una icona di warning, ciò avviene perché non esiste ancora una predizione per il classificatore "Test" su questa citazione.

Dopo aver aggiunto un nuovo classificatore nei file JSON e aver riavviato l'applicazione, sul database viene aggiunta una nuova colonna `prediction` che serve per contenere la predizione del nuovo classificatore su ogni citazione. In questo caso la entità `Citation` sul database viene aggiornata come mostrato nella figura 4.4.



Attenzione: per aggiungere con correttamente un classificatore è necessario che il valore del campo `prediction` finisca sempre per "_prediction"

4.5 Meccanismo di "agreement" tra classificatori

In questa sezione vediamo come fa l'applicazione a ricostruire un *mapping* e il livello di *agreement* tra i classificatori.

L'applicazione, per ogni citazione, riesce a costruire a interfaccia il *mapping* dei classificatori utilizzando il file JSON 4.4. Possiamo notare che ogni intento ha come valore associato un intero, questo rappresenta il suo peso e graficamente rappresenta una riga di una flexbox. Ciclando tutti gli intenti del classificatore riesce a ricostruire la tabella.

Algoritmo di agreement
In questo paragrafo descriviamo come la parte backend del CIT riesce a calcolare il livello di agreement tra le etichette predette da più classificatori su una citazione.

Quando l'utente invia la richiesta di caricamento delle citazioni, il backend le filtra e su ognuna di queste applica il metodo *getAgreement()* che aggiunge delle informazioni necessarie da fornire al frontend tra cui il livello di agreement.
Prendiamo una citazione di esempio:

```
citation = {
    creationTime: "2022-12-04T16:14:52.000Z"
    importedBy: "root",
    anchor_sentence: "test",
    biblio_entry: "test",
    category: "default",
    citation_id: 44,
    "cited_doi": "https://doi.org/10.1109/aina.2014.13",
    "citing_doi": "http://dx.doi.org/10.1016/j.adhoc.2015.05.004"
    aclarc_prediction: "uses",
    scicite_prediction: "method"
```

```
    scar_prediction: "uses_method",
}
```

Il metodo raccoglie inizialmente le predizioni, in questo caso tre, e per ognuna di queste crea un oggetto:

```
{
    prediction: string,
    classifier :string,
    row_number: number[],
    weight: number
}
```

Vengono memorizzate delle informazioni: il nome del classificatore, la sua predizione e il peso associato.
In seguito vengono sommati i pesi degli intenti per quel classificatore fino all'intento predetto trovando così in quale riga si trova nella tabella. Questa informazione viene salvata nel campo "row_number".
Avendo questa informazione per tutte tre le predizioni è possibile calcolare il livello di agreement facendo un confronto 1:1 vedendo se una delle righe che occupa la predizione A è inclusa nelle righe che occupa la predizione B.

1. uses(ACL-ARC) - method(SciCite) = AGREEMENT
2. method(SciCite) - uses_method(SCAR) = AGREEMENT
3. uses(ACL-ARC) - uses_method(SciCite) = AGREEMENT

In questo caso il livello di agreement è del 100%.

<table>
<tr><th>ACL-ARC</th><th>SciCite</th><th>SCAR</th></tr>
<tr><td>Extends</td><td rowspan="4">Background</td><td>Extends</td></tr>
<tr><td>Background</td><td rowspan="3">Cites</td></tr>
<tr><td>Motivation</td></tr>
<tr><td>Future</td></tr>
<tr><td>Compare</td><td>Compare result</td></tr>
<tr><td rowspan="2">Uses</td><td rowspan="2">Method</td><td>Uses method</td></tr>
<tr><td>Uses data from</td></tr>
</table>

Oltre al livello di agreement viene anche identificata la predizione considerata più accurata, ovvero quella che occupa meno righe di tutte. Per questa citazione la predizione suggerita è "uses_method" del classificatore "SCAR".
Come informazione aggiuntiva viene anche restituito il peso maggiore tra le etichette predette ovvero 2.

Tutte queste informazioni vengono quindi inserite all'interno della citazione in un nuovo campo chiamato "agreement". Seguendo l'esempio, mostro qui di seguito l'oggetto finale che viene restituito al frontend:

```
citation = {
    creationTime: "2022-12-04T16:14:52.000Z"
    importedBy: "root",
    anchor_sentence: "test",
    biblio_entry: "test",
    category: "default",
    citation_id: 44,
    "cited_doi": "https://doi.org/10.1109/aina.2014.13",
    "citing_doi": "http://dx.doi.org/10.1016/j.adhoc.2015.05.004"
    aclarc_prediction: "uses",
    scicite_prediction: "method"
    scar_prediction: "uses_method",
    "agreement": {
        "level": 1,
        "prediction": "uses_method",
        "classifier": "SCAR",
        "biggestRowSize": 2
    }
}
```

4.6 Updater Process

Fino a questo punto non abbiamo ancora visto come la nostra applicazione si collega e utilizza un classificatore di citazioni per ottenere delle predizioni.
Abbiamo dato per scontato che le citazioni avessero già una predizione per un classificatore, ma in realtà esiste un meccanismo che si occupa di comunicare con il modello per ottenere queste informazioni.
In questa sezione vediamo in pratica come un classificatore può esprimere una funzione su una determinata citazione all'interno della applicazione.

L'updater process è un processo separato sviluppato in codice javascript che utilizza l'ambiente di esecuzione fornito da Node.js. Il processo si interfaccia con il server dell'applicazione CIT e comunica tramite un apposita chiamata REST per raccogliere le citazioni che non hanno ancora una predizione. In seguito il processo cicla tutte le citazioni e cerca di ottenre le predizioni mancanti richiamando i classificatori di competenza, dopodichè le restituisce al server.

Sequenza di operazioni dell'Updater Process

Il flusso del processo si può riassumere in 4 passaggi:

STEP 1
Esegue inizialmente una chiamata REST al server per ottenere tutte le citazioni da processare. L'updater process non conosce la struttura del database ma il è server che tramite i file JSON dei classificatori riesce a costruire la query per ottenere le citazioni che hanno almeno uno dei campi "prediction" nullo.
Se abbiamo configurati i 3 classificatori ACL-ARC, SciSite e SCAR allora la query sarà:

```
SELECT *
FROM citation_environment.citation
WHERE (aclarc_prediction IS null
       OR scicite_prediction IS null
       OR scar_prediction IS NULL);
```

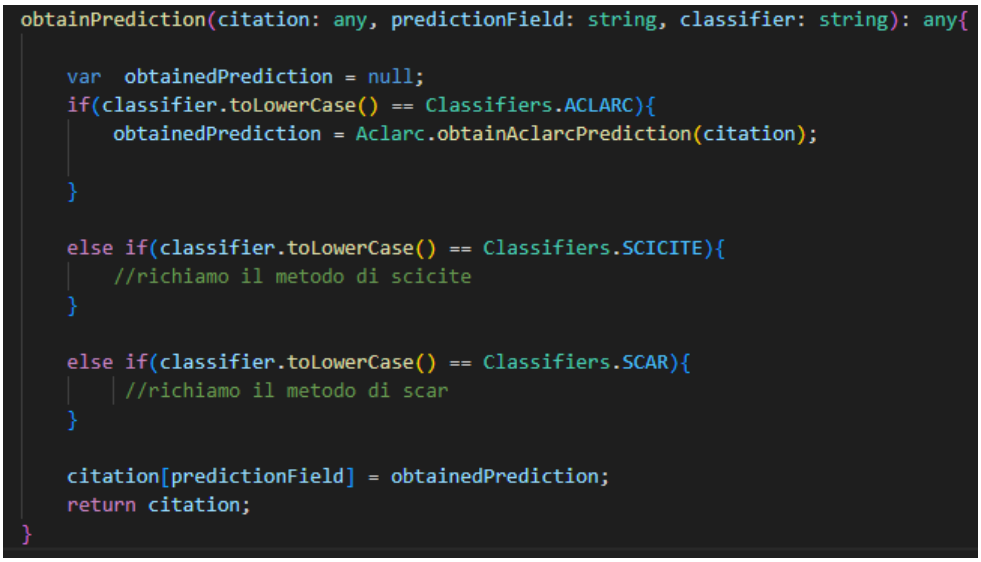
Le citazioni ottenute vengono poi raccolte dal processo in un array.

STEP 2
Il processo prende una citazione dall'array, per esempio:

```
citation = {
    creationTime: "2022-12-04T16:14:52.000Z"
    importedBy: "root",
    anchor_sentence: "test",
    biblio_entry: "test",
    category: "default",
    citation_id: 44,
    "cited_doi": "https://doi.org/10.1109/aina.2014.13",
    "citing_doi": "http://dx.doi.org/10.1016/j.adhoc.2015.05.004"
    aclarc_prediction: null,
    scicite_prediction: "method"
    scar_prediction: null",
}
```

Identifica quali campi sono nulli tra quelli che terminano per "_prediction", in questo caso "aclarc_prediction" e scar_prediction". Per ogni campo crea una coppia <campo, classificatore>. Da "scar_prediction" ricava quindi la coppia <scar_prediction, scar>.

STEP 3
Per ogni coppia campo-classificatore creata viene richiamato un "dispatcher".
Mostro di seguito il metodo che rappresenta il "dispatcher":



```
obtainPrediction(citation: any, predictionField: string, classifier: string): any{

    var  obtainedPrediction = null;
    if(classifier.toLowerCase() == Classifiers.ACLARC){
        obtainedPrediction = Aclarc.obtainAclarcPrediction(citation);

    }

    else if(classifier.toLowerCase() == Classifiers.SCICITE){
        //richiamo il metodo di scicite
    }

    else if(classifier.toLowerCase() == Classifiers.SCAR){
        //richiamo il metodo di scar
    }

    citation[predictionField] = obtainedPrediction;
    return citation;
}
```

Il metodo riceve in input la citazione, il campo della predizione e il nome del classificatore. Eseguendo una serie di controlli viene triggerato il giusto classificatore in modo da ottenere da esso una funzione citazionale. In questo sostituiamo il campo precedente nullo con la predizione ottenuta.

STEP 4
Esegue quindi gli step precedenti per tutte le citazioni presenti nella array. Infine restituisce tutte le citazioni modificate al server tramite una chiamata REST.

Adesso che abbiamo definiti questi passaggi, nella figura 4.4 diamo una rappresentazione grafica della struttura dell'Updater Process

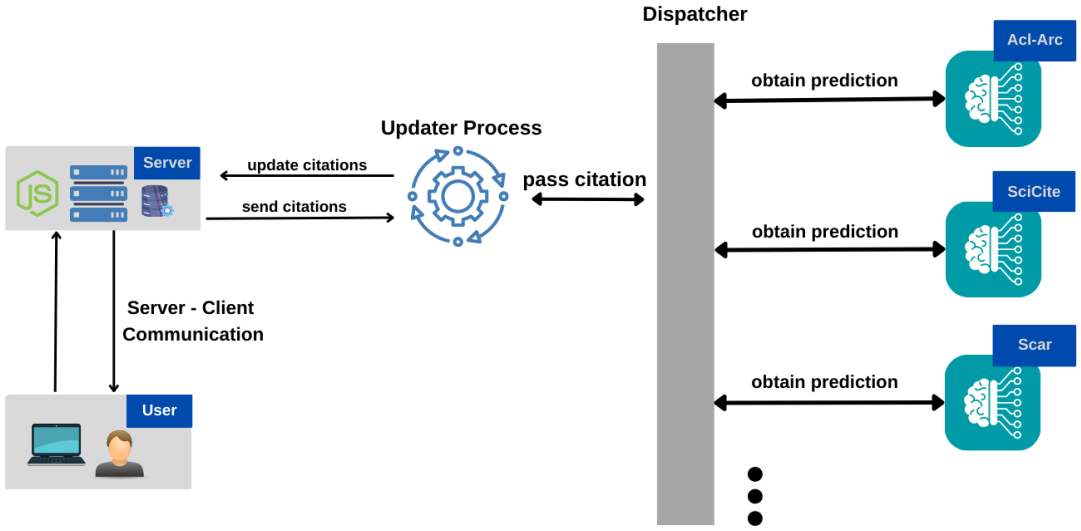


Figura 4.4: Panoramica dell'Updater Process

Capitolo 5

Conclusioni

Questa tesi nasce per dare un contributo all'analisi delle funzioni citazionali. Questa area di ricerca è nata dalla necessità di studiare le diverse modalità in cui le citazioni vengono utilizzate all'interno dei documenti accademici. Differenziando la funzione che svolge una citazione si riesce a dare un'indicazione sulla qualità e l'impatto che un articolo può avere in una specifica disciplina. Fornisce quindi un ulteriore criterio per evidenziare gli articoli e autori più autorevoli e il tipo di contributo che offrono. Sono stati proposti più modelli di machine learning e classificatori che identificano e categorizzano automaticamente le citazioni in un testo in base alle funzioni che svolgono. Questi classificatori vengono nella maggior parte dei casi progettati in base ad un singolo *task* considerando una determinata disciplina. Per questi motivi le etichette che propongono sono spesso differenti e i risultati ottenuti da più ricerche sono difficili da confrontare. I classificatori che si basano su modelli di apprendimento automatico hanno la necessità di utilizzare molti dati annotati e più questi sono accurati più i classificatori fanno predizioni migliori. Lo stesso processo di raccogliere dati annotati è lungo e dispendioso.
Per proseguire nella ricerca abbiamo quindi realizzato il CIT, un'applicazione web che fornisce un ambiente di annotazione di citazioni per la produzioni di dati accurati che verranno utilizzati dai classificatori attuali per migliorare le loro predizioni oppure per crearne di nuovi.
Il CIT offre i due fondamentali meccanismi di *mapping* e *agreement* per confrontare i risultati di più classificatori, permette di aiutare l'utente a scegliere la funzione citazionale migliore e in questo modo raccoglie dati annotati accurati.
Un punto di forza è che all'utente è data la possibilità di utilizzare dei criteri per filtrare le citazioni, questo dà flessibilità all'applicazione e permette all'annotatore o al gruppo di annotatori di diversificare i casi d'uso per i propri obiettivi.
Il CIT presenta un'interfaccia grafica semplice e intuitiva, grazie a ciò un utente può annotare velocemente e in modo accurato.
Nel caso fosse utilizzata da molti annotatori potrebbe diventare un punto di riferimento per l'accumulazione di citazioni, annotazioni e predizioni, e quindi fornire una fonte utile per gli studi delle funzioni citazionali.

Bibliografia

[1] H. D. White, Citation Analysis and Discourse Analysis Revisited, Applied Linguistics 25 (1) (2004) 89–116. arXiv:https://academic.oup.com/applij/article-pdf/25/1/89/348943/250089.pdf, doi:10.1093/applin/25.1.89.
URL https://doi.org/10.1093/applin/25.1.89

[2] J. E. Hirsch, An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences 102 (46) (2005) 16569–16572. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0507655102, doi:10.1073/pnas.0507655102.
URL https://www.pnas.org/doi/abs/10.1073/pnas.0507655102

[3] S. Bonzi, Characteristics of a literature as predictors of relatedness between cited and citing works, Journal of the American Society for Information Science 33 (4) (1982) 208, ultimo aggiornamento - 2013-02-24.
URL https://www-proquest-com.ezproxy.unibo.it/scholarly-journals/characteristics-literature-as-predictors/docview/1301249264/se-2

[4] J. M. Ziman, Public Knowledge: An Essay Concerning the Social Dimension of Science, London: Cambridge University Press, 1968.

[5] Monde: essai d'universalisme: connaissance du monde, sentiment du monde, action organisée et plan du monde, 1970.

[6] M. Garzone, R. E. Mercer, Towards an automated citation classifier, in: H. J. Hamilton (Ed.), Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 337–346.

[7] S. B. Pham, A. Hoffmann, A new approach for scientific citation classification using cue phrases, in: T. T. D. Gedeon, L. C. C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 759–771.

[8] S. Teufel, A. Siddharthan, D. Tidhar, An annotation scheme for citation function, in: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics, pp. 80–87.
URL https://aclanthology.org/W06-1312

[9] A. Abu-Jbara, J. Ezra, D. Radev, Purpose and polarity of citation: Towards NLP-based bibliometrics, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta,

Georgia, 2013, pp. 596–606.
URL https://aclanthology.org/N13-1067

[10] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, D. Jurafsky, Measuring the evolution of a scientific field through citation frames, Transactions of the Association for Computational Linguistics 6 (2018) 391–406. doi:10.1162/tacl_a_00028.
URL https://aclanthology.org/Q18-1028

[11] S. Bird, R. Dale, B. Dorr, B. R. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, Y. F. Tan, The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics, in: International Conference on Language Resources and Evaluation, 2008.

[12] A. Cohan, W. Ammar, M. van Zuylen, F. Cady, Structural scaffolds for citation intent classification in scientific publications, CoRR abs/1904.01608 (2019). arXiv:1904.01608.
URL http://arxiv.org/abs/1904.01608

[13] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780. arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf, doi:10.1162/neco.1997.9.8.1735.
URL https://doi.org/10.1162/neco.1997.9.8.1735

[14] P. Ciancarini, A. Di Iorio, A. G. Nuzzolese, S. Peroni, F. Vitali, Evaluating citation functions in cito: Cognitive issues, in: V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (Eds.), The Semantic Web: Trends and Challenges, Springer International Publishing, Cham, 2014, pp. 580–594.

[15] S. N. Kunnath, D. Herrmannova, D. Pride, P. Knoth, A meta-analysis of semantic classification of citations 2 (4) 1170–1215. doi:10.1162/qss_a_00159.
URL https://direct.mit.edu/qss/article/2/4/1170/107610/A-meta-analysis-of-semantic-classification-of

[16] M. Roman, A. Shahid, S. Khan, A. Koubaa, L. Yu, Citation intent classification using word embedding 9 9982–9995. doi:10.1109/ACCESS.2021.3050547.
URL https://ieeexplore.ieee.org/document/9319154/

[17] X. Zhu, P. Turney, D. Lemire, A. Vellino, Measuring academic influence: Not all citations are equal 66 (2) 408–427. doi:10.1002/asi.23179.
URL https://onlinelibrary.wiley.com/doi/10.1002/asi.23179

[18] The use of citations in literary research: A preliminary classification of citation functions 49 (4) 399–414. doi:10.1086/600930.
URL https://www.journals.uchicago.edu/doi/10.1086/600930

[19] M. J. Moravcsik, P. Murugesan, Some results on the function and quality of citations 5 (1) 86–92, publisher: Sage Publications, Ltd.
URL https://www.jstor.org/stable/284557

[20] S. Peroni, D. Shotton, FaBiO and CiTO: Ontologies for describing bibliographic resources and citations 17 33–43. doi:10.1016/j.websem.2012.08.001.
URL https://www.sciencedirect.com/science/article/pii/S1570826812000790

[21] S. Iqbal, S.-U. Hassan, N. R. Aljohani, S. Alelyani, R. Nawaz, L. Bornmann, A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies 126 (8) 6551–6599. doi:10.1007/s11192-021-04055-1.
URL https://doi.org/10.1007/s11192-021-04055-1

Ringraziamenti

Voglio ringraziare tutta la mia famiglia che mi ha sempre supportato e non mi ha mai fatto mancare nulla. In particolare ringrazio Silvia che in questi tre anni di università mi ha sempre aiutato e sopportato.

Ringrazio i miei amici e compagni di corso Stefano, Mattia e Gabriele. Mi sono trovato benissimo con voi, siamo riusciti a portare avanti e concludere tanti progetti insieme. Abbiamo riso e scherzato e abbiamo passato dei bei momenti. Spero in futuro di collaborare ancora con voi.

Ringrazio Tommaso, Divo, Dejvi e Giulia che mi sono stati molto vicini e che sono da sempre una sicurezza.